

# Modelo de Previsão de Novos Casos de COVID-19 e Mortes Relacionadas à Doença



29/05/2020

# Pós-Graduação "Lato Sensu" Especialização em Análise de Big Data

**Nome do Aluno:**

Gustavo de Carvalho Ferreira e Vieira

**Coordenadores:**

Prof.<sup>a</sup> Dr.<sup>a</sup> Alessandra de Álvila Montini

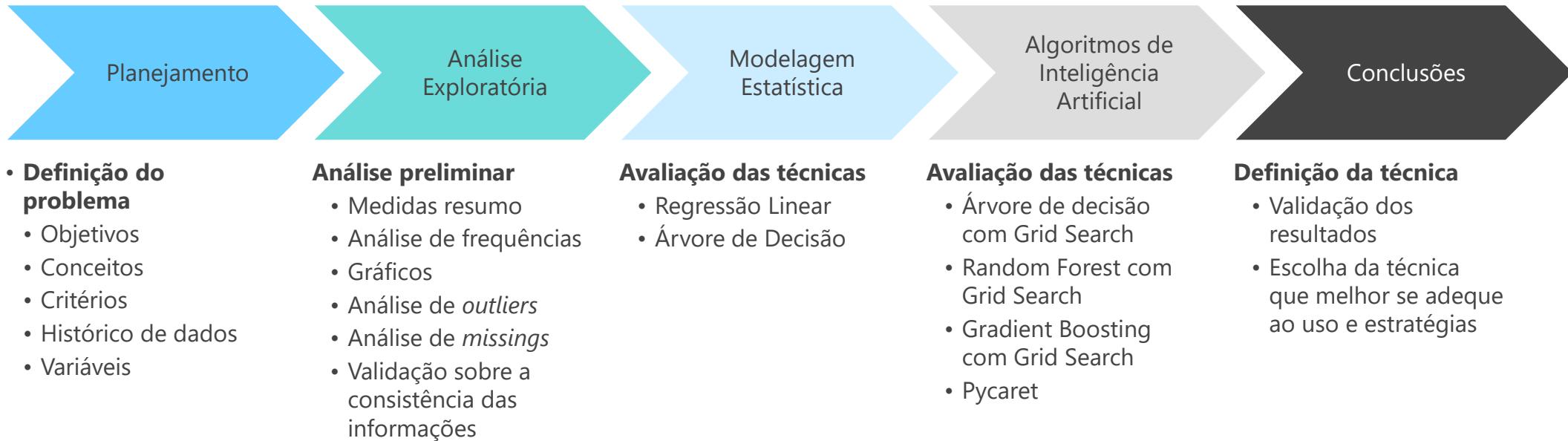
Prof. Dr. Adolpho Walter Pimazoni Canton

# Agenda

- 1. Objetivo do Trabalho
- 2. Contextualização do Problema
- 3. Base de Dados
  - i. Bases originais
  - ii. Transformações das Bases de Dados
  - iii. Principais Variáveis Após Transformações
  - iv. Processo de Redução de Variáveis
- 4. Análise Exploratória de Dados
- 5. Modelagem com Estatística Tradicional
- 6. Modelagem com Inteligência Artificial
- 7. Conclusões
- 8. Sugestão para Trabalhos Futuros

# Metodologia de Análise de Dados

4

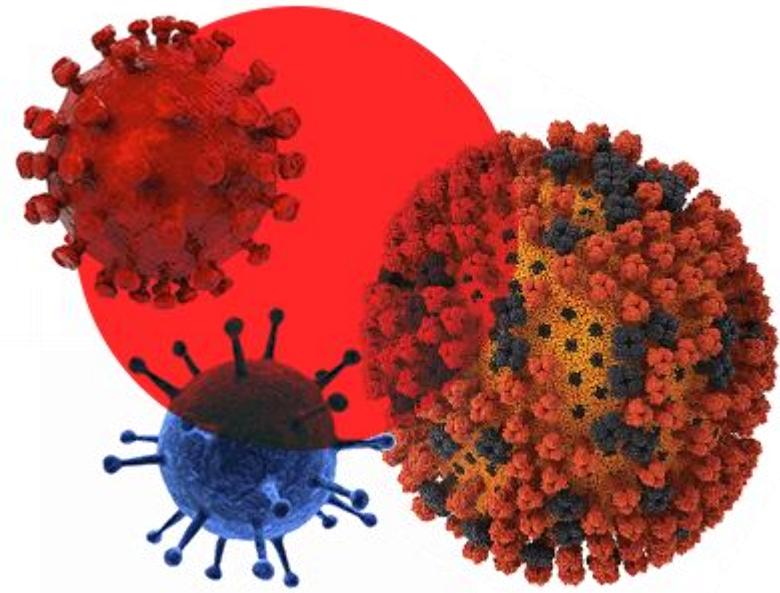


# Metodologia de Análise de Dados

5



# 1. Objetivo do Trabalho



## Modelo de Previsão de Novos Casos de COVID-19 e Mortes Relacionadas à Doença

Inicialmente, o objetivo deste trabalho era **prever** a curva de **casos de COVID-19** na população do município de São Paulo, do estado de São Paulo e do Brasil e de **mortes** relacionadas à doença, a fim de municiar os entes governamentais de informações para **tomada de decisão na saúde pública** em meio à pandemia. Durante a análise exploratória, percebemos que a qualidade dos dados de vários estados brasileiros eram questionáveis, por isso, a partir da segunda entrega, reformulamos o objetivo para nos concentrar apenas nos **municípios do estado de São Paulo**, que além de terem dados mais confiáveis, também traziam um dado adicional: o índice de isolamento social.

A previsão será realizada por meio da análise do banco de dados histórico e uso de **modelos estatísticos** e **algoritmos de Inteligência Artificial**, que selecionarão as características mais relevantes que explicam os eventos de transmissão e mortes.

Desta forma, os governos municipais e estaduais poderão traçar **estratégias de prevenção** (distanciamento social, lockdown, retomada das atividades econômicas), **de criação de leitos de UTIs públicos** e **de aluguel de leitos de UTI privados**, de acordo com a fase da pandemia.

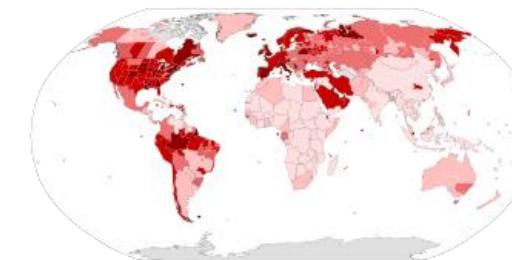


## 2. Contextualização do Problema

A pandemia de COVID-19 teve início na cidade de Wuhan, na China, em dezembro de 2019.

Causada pelo SARS-CoV-2, uma variação do coronavírus que causou a Síndrome Respiratória Aguda Grave (SARS) na epidemia de 2002, a doença rapidamente se espalhou pelo mundo, com a Organização Mundial de Saúde (OMS) declarando em 30 de janeiro de 2020 que a epidemia era uma Emergência de Saúde Pública de Âmbito Internacional. Com a escalada de casos, em 11 de março de 2020 a OMS declarou estarmos vivendo uma pandemia. O primeiro caso registrado oficialmente no Brasil aconteceu na cidade de São Paulo, em 26 de fevereiro de 2020.

Com **altos índices de contágio** através de gotículas de saliva, a doença ataca primariamente os pulmões. Em **casos graves**, o paciente tem queda drástica de oxigenação no sangue e **necessita internação em leito de Unidade de Terapia Intensiva (UTI)** e entubação para receber ventilação mecânica. Como há **oferta limitada de leitos** de UTI, é necessário **estimar a quantidade de novos casos de COVID-19 para direcionar políticas** de enfrentamento **que evitem o colapso do sistema de saúde**, como isolamento social, criação de novos leitos de UTI e requisição de leitos de UTI privados pelo poder público.

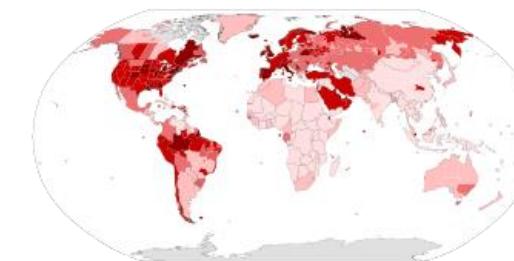


## 2. Contextualização do Problema

Há que se levar em consideração as dificuldades de modelagem, a saber:

- É consenso entre os pesquisadores do tema de que há imensa subnotificação dos casos de COVID-19 no Brasil, devido a falta de testes em massa, vontade política e falta de bases de dados únicas e universais.
- Os números de casos e mortes confirmados disponibilizados pelas secretarias de saúde estaduais e municipais, que compõem a base da Lagom Data, sofrem modificações históricas constantes, conforme os resultados dos testes são liberados, com casos e mortes passadas sendo agregados aos dados históricos. Além disso, alguns estados alteraram o critério de identificação do município durante o período de coleta de dados, de local de diagnóstico para local de residência do paciente.
- Não há padronização das informações disponíveis ao público, e em casos como o histórico de taxa de isolamento e o histórico de ocupação dos leitos de UTI, variáveis importantes para prever a aceleração da epidemia e o colapso do sistema de saúde, as informações não estão disponíveis para a grande maioria dos municípios.

Portanto, o estudo a seguir não tem a intenção de ser um modelo definitivo e com altíssima precisão, mas um norte em termos de ordens de grandeza.



### 3. Bases de Dados

Para esse estudo, serão utilizadas as seguintes bases de dados:

- Base de dados compilados sobre COVID-19 pela **Lagom Data** (<https://www.lagomdata.com.br/coronavirus>), que por sua vez utilizou como fontes as Secretarias Estaduais de Saúde, o Ministério da Saúde e o IBGE.
  - A base é composta de uma série histórica de casos confirmados da doença e mortes em decorrência dela, com 21 variáveis e 55.869 observações, compreendendo o período de 26/02/2020 a 07/05/2020.
- Bases de dados de quantidade de leitos de internação hospitalar SUS e Não SUS por município e por tipo de leito, provenientes do Cadastro Nacional de Estabelecimentos de Saúde (CNES) do governo federal (<http://tabnet.datasus.gov.br/cgi/deftohtm.exe?cnes/cnv/leiintbr.def>), extraídas em fevereiro, março e abril de 2020.
- Base de dados de índice de isolamento social dos municípios do estado de São Paulo, proveniente da Secretaria de Saúde do Estado de São Paulo (<https://www.saopaulo.sp.gov.br/coronavirus/isolamento/>), extraída em 26/05/2020.



### 3.i. Bases Originais

BASE PRINCIPAL | Histórico coronavírus

10

Descrição da base disponibilizada pela Lagom Data:

Quantidade de registros	Quantidade de variáveis	Nomes das variáveis	Período analisado
<ul style="list-style-type: none"><li>• 55.869</li></ul>	<ul style="list-style-type: none"><li>• 21</li></ul>	<ul style="list-style-type: none"><li>• 'data', 'cidade', 'uf', 'confirmados', 'mortos', 'cod7d', 'munuf', 'habitantes', 'mesorregiao', 'confirmados_100mil', 'mortos_milhao', 'lat', 'lon', 'amazonia', 'fronteira', 'capital', 'litoral', 'semiariano', 'papel', 'idhm_2010', 'faixa_pop'</li></ul>	<ul style="list-style-type: none"><li>• 26/02/2020 a 07/05/2020</li></ul>



### 3.i. Bases Originais

BASES AUXILIARES | Leitos de Internação SUS e Não SUS Fev, Mar e Abr 2020

11

Descrição das bases auxiliares extraídas do CNES:

Quantidade de tabelas	Quantidade de registros	Quantidade de variáveis	Nomes das variáveis	Período analisado
<ul style="list-style-type: none"><li>• 6:<ul style="list-style-type: none"><li>• Leitos de Internação SUS Fev 2020</li><li>• Leitos de Internação Não SUS Fev 2020</li><li>• Leitos de Internação SUS Mar 2020</li><li>• Leitos de Internação Não SUS Mar 2020</li><li>• Leitos de Internação SUS Abr 2020</li><li>• Leitos de Internação Não SUS Abr 2020</li></ul></li></ul>	<ul style="list-style-type: none"><li>• 5.596 por tabela</li></ul>	<ul style="list-style-type: none"><li>• 8 por tabela</li></ul>	<ul style="list-style-type: none"><li>• 'Município', 'Cirúrgicos', 'Clínicos', 'Obstétrico', 'Pediátrico', 'Outras Especialidades', 'Hospital/DIA', 'Total'</li></ul>	<ul style="list-style-type: none"><li>• Fevereiro, Março e Abril de 2020</li></ul>



### 3.i. Bases Originais

BASES AUXILIARES | Índice de Isolamento Social

12

Descrição da base auxiliar extraída da Secretaria de Saúde do Governo do Estado de São Paulo:

Quantidade de tabelas	Quantidade de registros	Quantidade de variáveis	Nomes das variáveis	Período analisado
<ul style="list-style-type: none"><li>• 1:<ul style="list-style-type: none"><li>• Índice de isolamento por data e por município do estado de SP</li></ul></li></ul>	<ul style="list-style-type: none"><li>• 7.770</li></ul>	<ul style="list-style-type: none"><li>• 4</li></ul>	<ul style="list-style-type: none"><li>• 'Município1', 'UF1', 'Data', 'Índice de Isolamento'</li></ul>	<ul style="list-style-type: none"><li>• 5 de março de 2020 a 7 de maio de 2020.</li></ul>



### 3.ii. Transformações das Bases de Dados

13



### 3.ii. Transformações das Bases de Dados

14

Durante as fases de preparação das bases e da análise exploratória, foram detectados problemas com os dados registrados, que exigiram transformações extras:



#### Transformações

- **Nomes de municípios versus código de municípios:** havia mais nomes do que códigos, devido a erros de grafia. Utilizamos uma nova base auxiliar do IBGE, com as grafias oficiais para corrigir o problema e acrescentar a informação da microrregião.
- **Casos acumulados e mortes acumuladas negativas:** consultas aos boletins expedidos pelas Secretarias de Saúde levaram à detecção de falhas na aquisição dos dados pela Lagom Data e retificações das próprias Secretarias de Saúde. Com isso, reconstituímos os históricos.
- **IDHM:** como o indicador é atualizado apenas no censo, havia municípios que foram criados após o censo de 2010, onde imputamos valor de IDHM = 0.
- **Casos diários e mortes diárias:** já eram esperados números negativos nesses 2 indicadores, devido às retificações de casos realizadas pelas secretarias de saúde, mas encontramos valores negativos muito expressivos em relação aos acumulados para serem simples retificações. A investigação comprovou falhas na aquisição dos dados pela Lagom Data, além de inconsistência de dados da Secretaria de Saúde do ES. Entre excluir ES da análise e inferir médias, optamos pela 3<sup>a</sup> via: restringir o escopo do trabalho e focar apenas no estado de SP, onde reconstituímos o histórico com base nos boletins da secretaria de saúde estadual.
- **Filtro de dados SP:** das 55869 linhas com dados de todo o Brasil, ficamos com 8795 linhas apenas dos municípios do estado de SP.



### 3.iii. Principais variáveis após transformações

Variáveis criadas | Variáveis originais

15



#### Variáveis IBGE

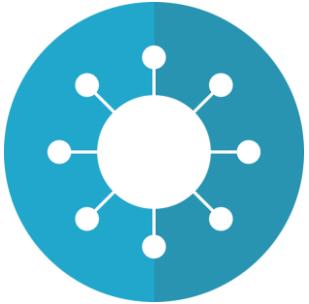
- Cidade: nome do município
- UF: nome do estado
- Cod7d: código de 7 dígitos do município
- Munuf: município-UF
- Habitantes: população estimada do município em 2019
- Mesorregião: agrupamento geográfico do IBGE
- Microrregião: agrupamento geográfico do IBGE
- Lat: latitude do município
- Lon: longitude do município
- Amazonia: indica se o município faz parte da Amazônia Legal
- Fronteira: indica se o município faz parte da fronteira com outros países
- Capital: indica se o município é capital de estado
- Litoral: indica se o município faz parte do litoral
- Semiarido: indica se o município faz parte do semiárido
- Papel: indica se o município é capital, interior ou da região metropolitana
- Idhm\_2010: índice de desenvolvimento humano municipal de acordo com último censo
- Faixa\_pop: indica em qual faixa de população o município está: maior de 100 mil, maior de 50 mil, maior de 10 mil ou menor de 10 mil



### 3.iii. Principais variáveis após transformações

Variáveis criadas | Variáveis originais

16



#### Variáveis COVID-19

- Data: dia, mês e ano em que caso de COVID-19 foi confirmado e/ou que morte por COVID-19 ocorreu
- Casos acumulados: número acumulado de casos confirmados de COVID-19 até a data
- Mortes acumuladas: número acumulado de mortos por COVID-19 confirmados até a data
- Casos por 100mil hab: indicador de casos confirmados acumulados até a data por 100 mil habitantes
- Mortes por milhão hab: indicador de mortes confirmadas acumuladas até a data por milhão de habitantes
- Índice de isolamento: indica percentual de pessoas do município que não se movimentou fora do raio da residência em determinado dia
- Casos diários: obtido através da subtração dos casos acumulados do dia anterior dos casos acumulados do dia
- Mortes diárias: obtido através da subtração das mortes acumuladas do dia anterior das mortes acumuladas do dia
- Mês: mês em que houve caso confirmado e/ou morte (criado para servir de índice ao cruzar com os dados das variáveis de leitos de UTI)
- Dias epidemiológicos: a quantidade de dias passados desde o primeiro caso, representando há quantos dias a cidade está na epidemia.



### 3.iii. Principais variáveis após transformações

Variáveis criadas | Variáveis originais

17



#### Variáveis Leitos de UTI

- Município: nome do município
- Cirúrgicos: quantidade de leitos de UTI cirúrgicos SUS e Não SUS
- Clínicos: quantidade de leitos de UTI clínicos SUS e Não SUS
- Obstétrico: quantidade de leitos de UTI obstétricos SUS e Não SUS
- Pediátrico: quantidade de leitos de UTI pediátricos SUS e Não SUS
- Outras Especialidades: quantidade de leitos de UTI de outras especialidades não citadas SUS e Não SUS
- Hospital/DIA: quantidade de leitos de UTI de assistência intermediária entre a internação e o atendimento ambulatorial, que requeiram a permanência do paciente na Unidade por um período máximo de 12 horas SUS e Não SUS
- Total: soma das quantidade de leitos de UTI de todos os tipos
- **Mês: mês em que os dados foram consolidados**



### 3.iv. Processo de Redução de Variáveis

18

#### Recebidas

66

- Base Principal: 21
- Bases Auxiliares de Leitos de UTI: 8 x 4
- Base Auxiliar Índice de Isolamento: 4
- Base Auxiliar de Municípios do IBGE: 9

#### Joins e drops de variáveis repetidas e não úteis

-28 = 38

#### • Base DadosSP

- Exemplos de variáveis consideradas não úteis: UF, amazonia, fronteira, semiárido.

#### Construídas na Análise Exploratória

+3 = 41

#### • Variáveis construídas:

- Zona geográfica (com base na latitude)
- Faixa meridional (com base na longitude)
- Dias epidemiológicos (quantidade de dias passados desde o primeiro caso, representando há quantos dias a cidade está na epidemia)

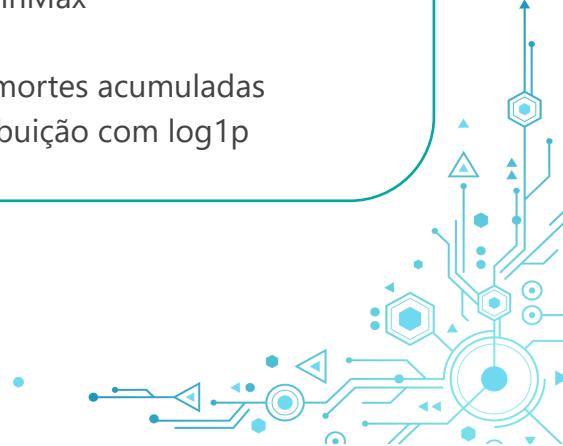
#### ABT

Explicativas: 59

Targets: 2

#### • Processos realizados:

- Explicativas (59):
  - Variáveis excluídas:
    - Exemplos de variáveis consideradas não úteis: código do município, derivadas de casos e mortes acumulados
  - Variáveis criadas:
    - Dia do ano, em substituição à data
    - Lags de 14 dias para casos e mortes acumuladas
    - Dummies para baixa cardinalidade
    - Label encoder para alta cardinalidade
    - Normalização com MinMax
  - Target (2)
    - casos acumulados e mortes acumuladas
    - Modificação da distribuição com log1p



# Analytic Base Table

	data	dias_epidemiológicos	dia_do_ano	mês	munuf	casos_acumulados	mortes_acumuladas	habitantes	lat	lon	idhm_2010
0	2020-02-26	1	57	2	São Paulo-SP	1	0	12252023	-23.54800	-46.63600	0.80500
1	2020-02-27	2	58	2	São Paulo-SP	1	0	12252023	-23.54800	-46.63600	0.80500
2	2020-02-28	3	59	2	São Paulo-SP	1	0	12252023	-23.54800	-46.63600	0.80500
3	2020-02-29	4	60	2	São Paulo-SP	1	0	12252023	-23.54800	-46.63600	0.80500
4	2020-03-01	5	61	3	São Paulo-SP	1	0	12252023	-23.54800	-46.63600	0.80500

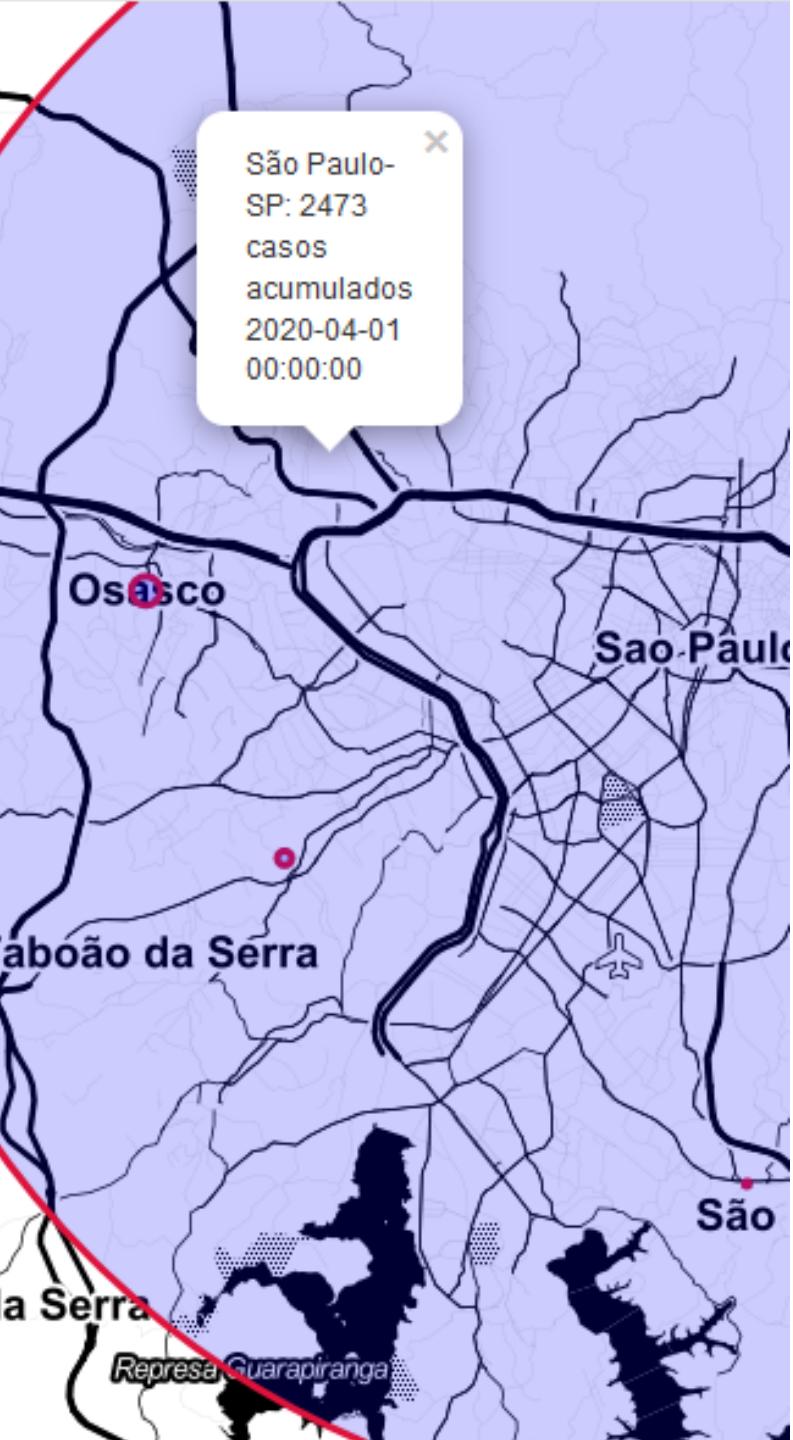
Nome_Mesorregião	Cirúrgicos_Não_SUS	Cínicos_Não_SUS	Obstétrico_Não_SUS	Pediátrico_Não_SUS	Outras_Especialidades_Não_SUS	HospitalDIA_Não_SUS
Metropolitana de São Paulo	4777	5331	1068	917	657	675
Metropolitana de São Paulo	4777	5331	1068	917	657	675
Metropolitana de São Paulo	4777	5331	1068	917	657	675
Metropolitana de São Paulo	4777	5331	1068	917	657	675
Metropolitana de São Paulo	4611	5258	1031	920	476	649

HospitalDIA_SUS	indice_isolamento	casos_acumulados_menos1d	casos_acumulados_menos2d	casos_acumulados_menos3d	casos_acumulados_menos4d	cas
789	0.53000	0	0	0	0	0
789	0.53000	1	0	0	0	0
789	0.53000	1	1	0	0	0
789	0.53000	1	1	1	0	0
789	0.53000	1	1	1	1	1

# Metodologia de Análise de Dados

20





## 4. Análise Exploratória de Dados

RAIO-X DA BASE | ESTRUTURA

### Tamanho do DataSet

8795 linhas  
43 colunas



### Valores Únicos

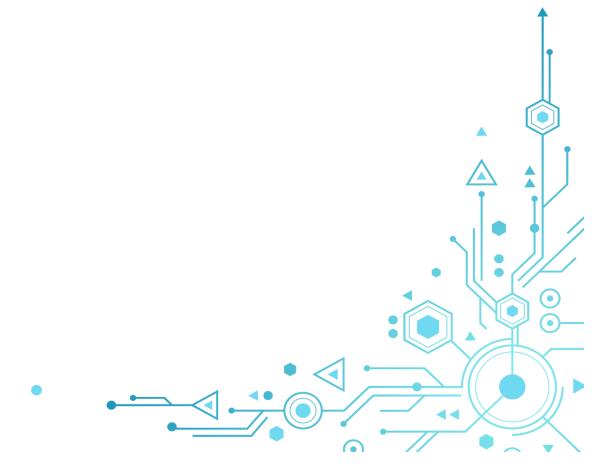
Com o escopo reduzido apenas ao estado de SP, não faz mais sentido levar todas as colunas para o modelo, pois algumas possuem o mesmo valor para todas as observações. Serão excluídas as colunas:

UF (SP)  
amazonia (Não)  
fronteira (Não)  
semiarido (Não)



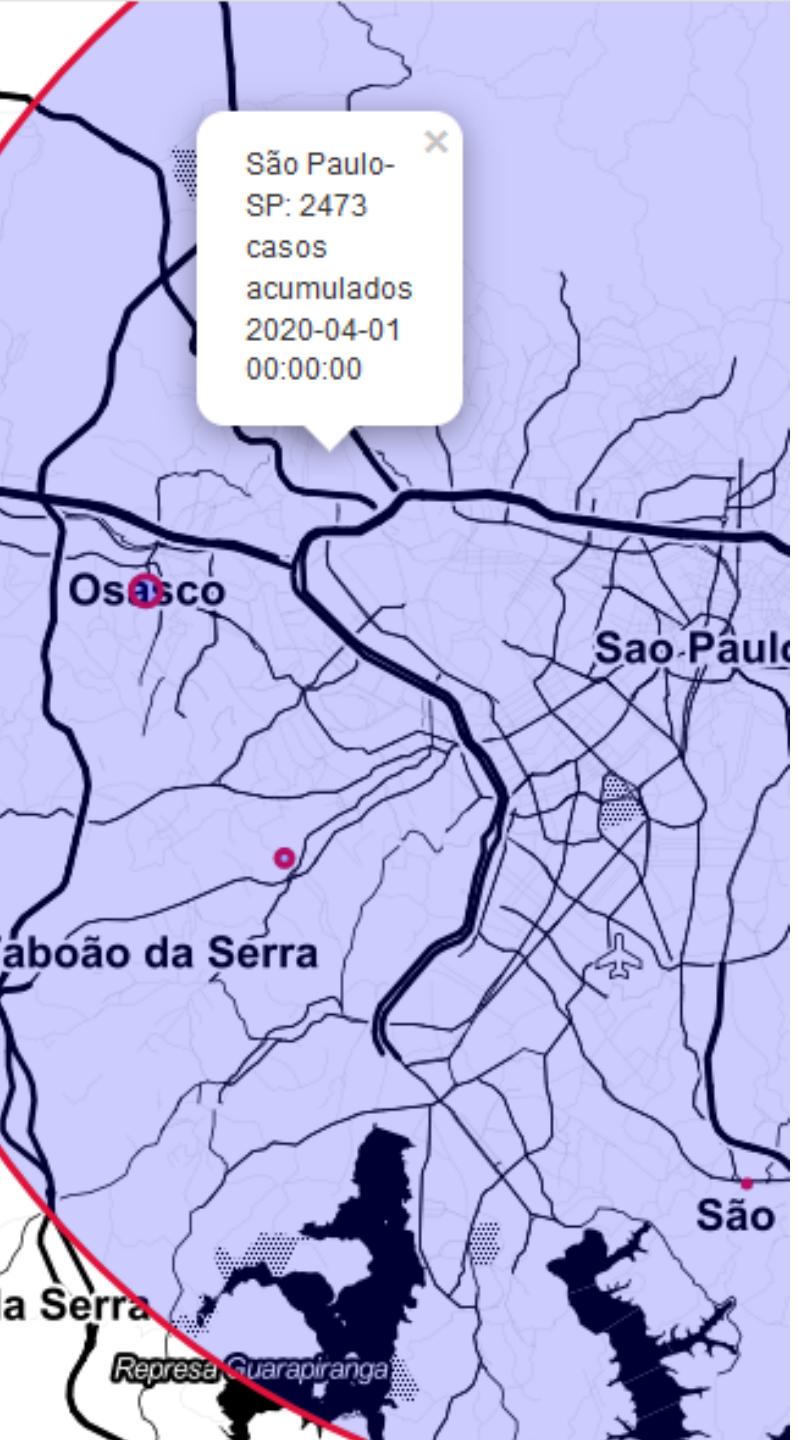
### Valores Faltantes

Como pouco mais da metade das cidades paulistas não possuem índice de isolamento, vamos considerar o valor da mediana do índice de isolamento (0.53) no lugar dos valores faltantes.



## 4. Análise Exploratória de Dados

RAIO-X DA BASE | NOVA VARIÁVEL



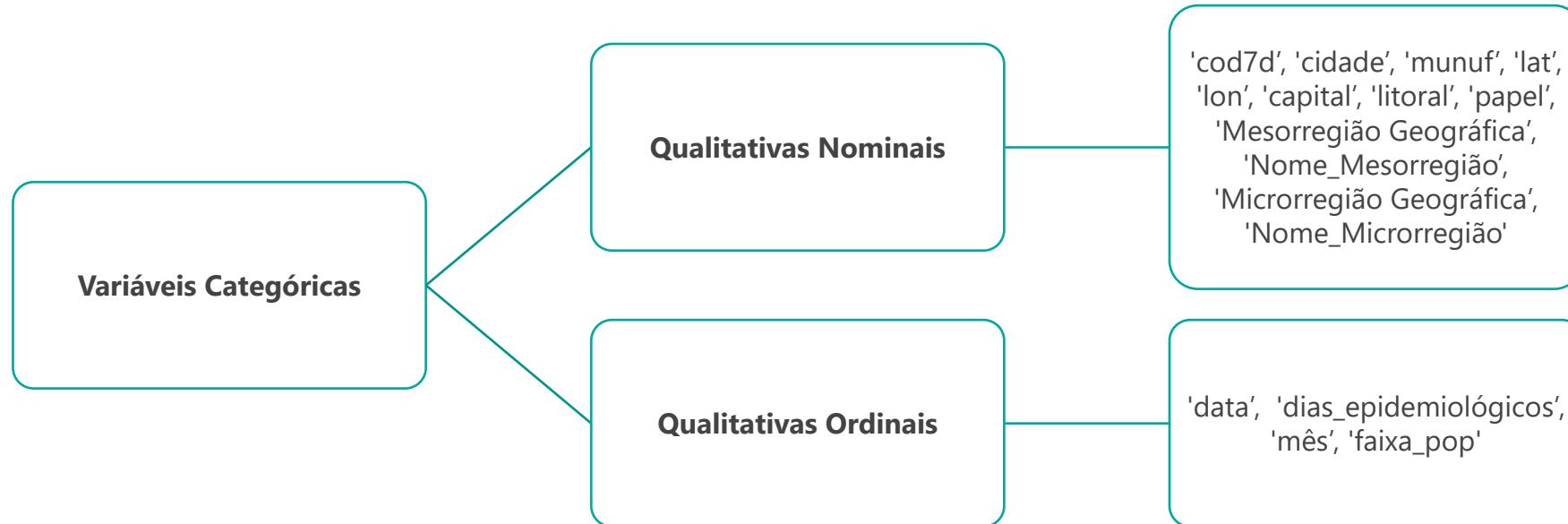
Dia  
epidemiológico

- É a quantidade de dias passados desde o primeiro caso, representando há quantos dias a cidade está na epidemia.
- Essa variável poderá ser utilizada na modelagem para cálculos de previsão de casos e mortes.
- Exemplo: no dia do ano 123, uma cidade pode estar no dia 5 da epidemia e outra cidade pode estar no dia 120 da epidemia: portanto, não se pode multiplicar o mesmo dia do ano por uma constante e esperar que dê resultados diferentes para as duas cidades. Mas se multiplicar o dia epidemiológico de cada uma pela constante, teremos um número de casos diferente e proporcional à fase em que cada cidade se encontra.

# 4. Análise Exploratória de Dados

RAIO-X DA BASE | CLASSIFICAÇÃO DAS VARIÁVEIS

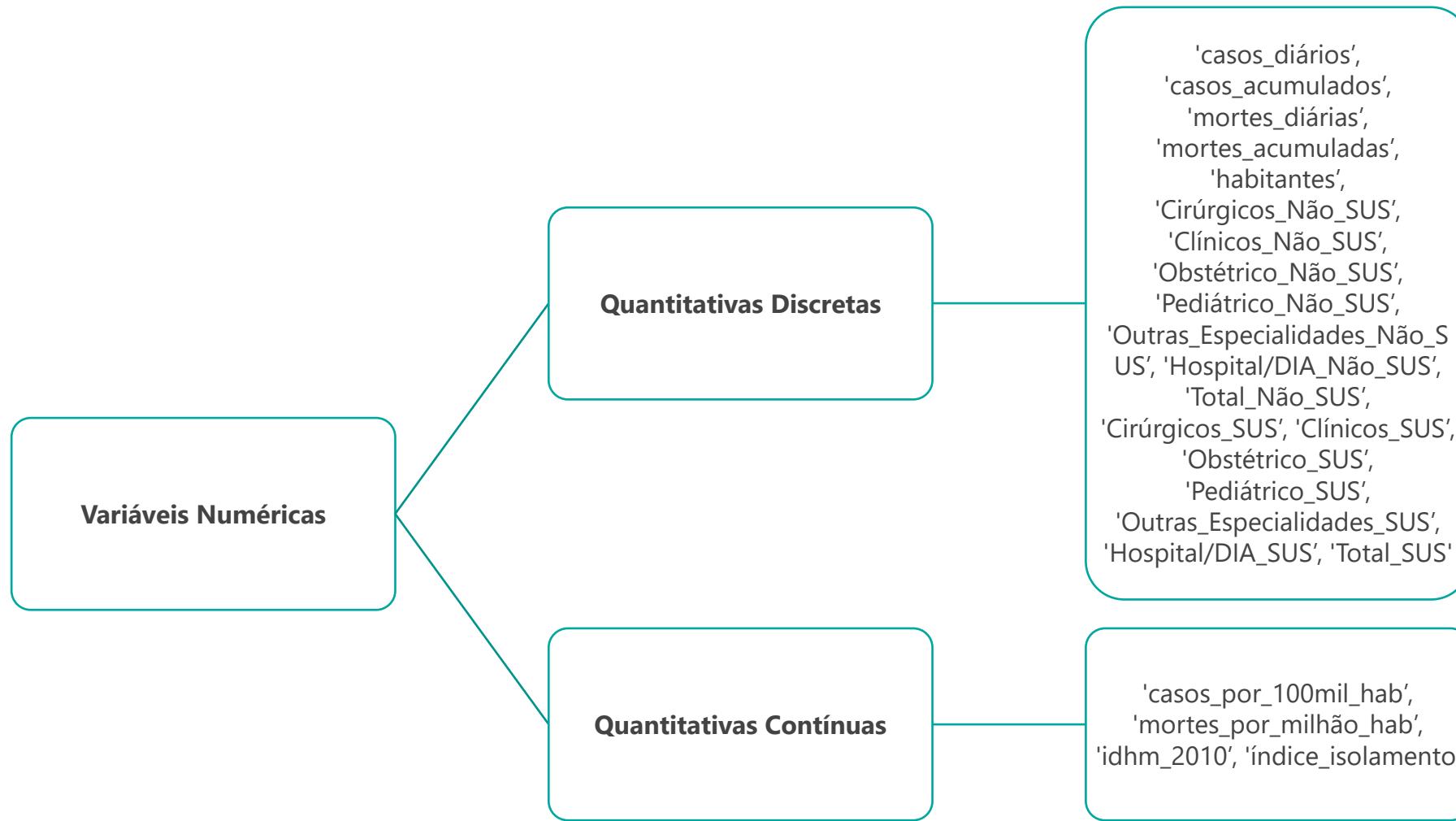
23



# 4. Análise Exploratória de Dados

RAIO-X DA BASE | CLASSIFICAÇÃO DAS VARIÁVEIS

24



## 4. Análise Exploratória de Dados

RAIO-X DA BASE | ANÁLISE UNIVARIADA



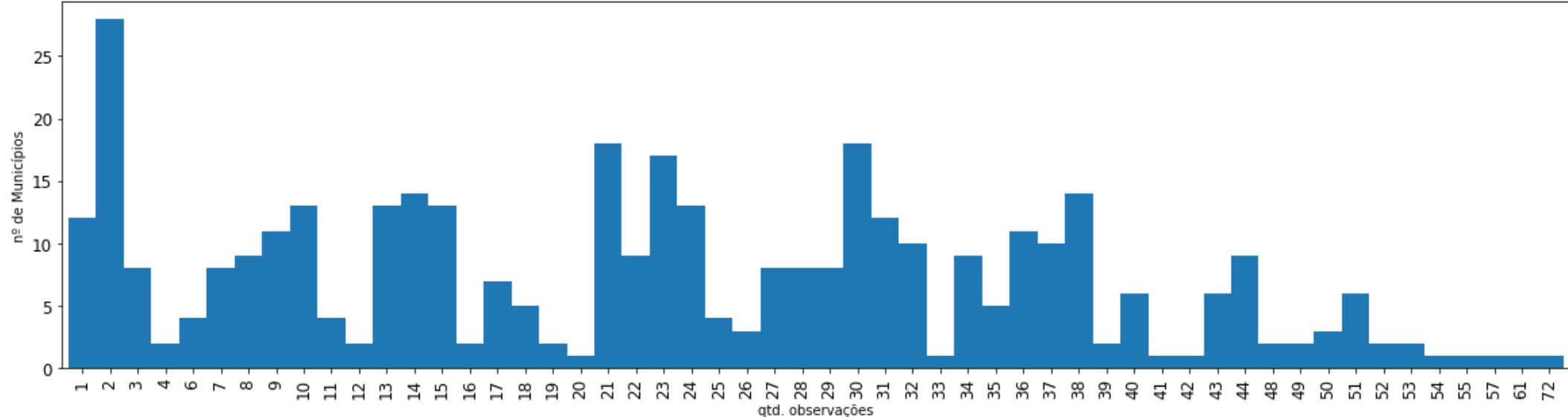
Detalhes das análises

25

'cod7d', 'cidade', 'munuf'

- São Paulo-SP é a única cidade que aparece com observações nas 72 datas por ter sido a 1ª cidade brasileira a confirmar um caso de COVID-19 (homem de 61 anos que voltou de uma viagem na Itália)
- A concentração de cidades com apenas 1 ou 2 observações acontece devido à transmissão sustentada: a doença não surge espontaneamente em um município, ela é levada ao município por uma pessoa infectada e a partir daí começa a se alastrar com o contágio das pessoas do município.
- 158 municípios paulistas têm até 20 observações.
- 82 municípios paulistas têm 45 observações ou mais.

Quantidade de Observações por Municípios do Estado de SP



# 4. Análise Exploratória de Dados

RAIO-X DA BASE | ANÁLISE UNIVARIADA

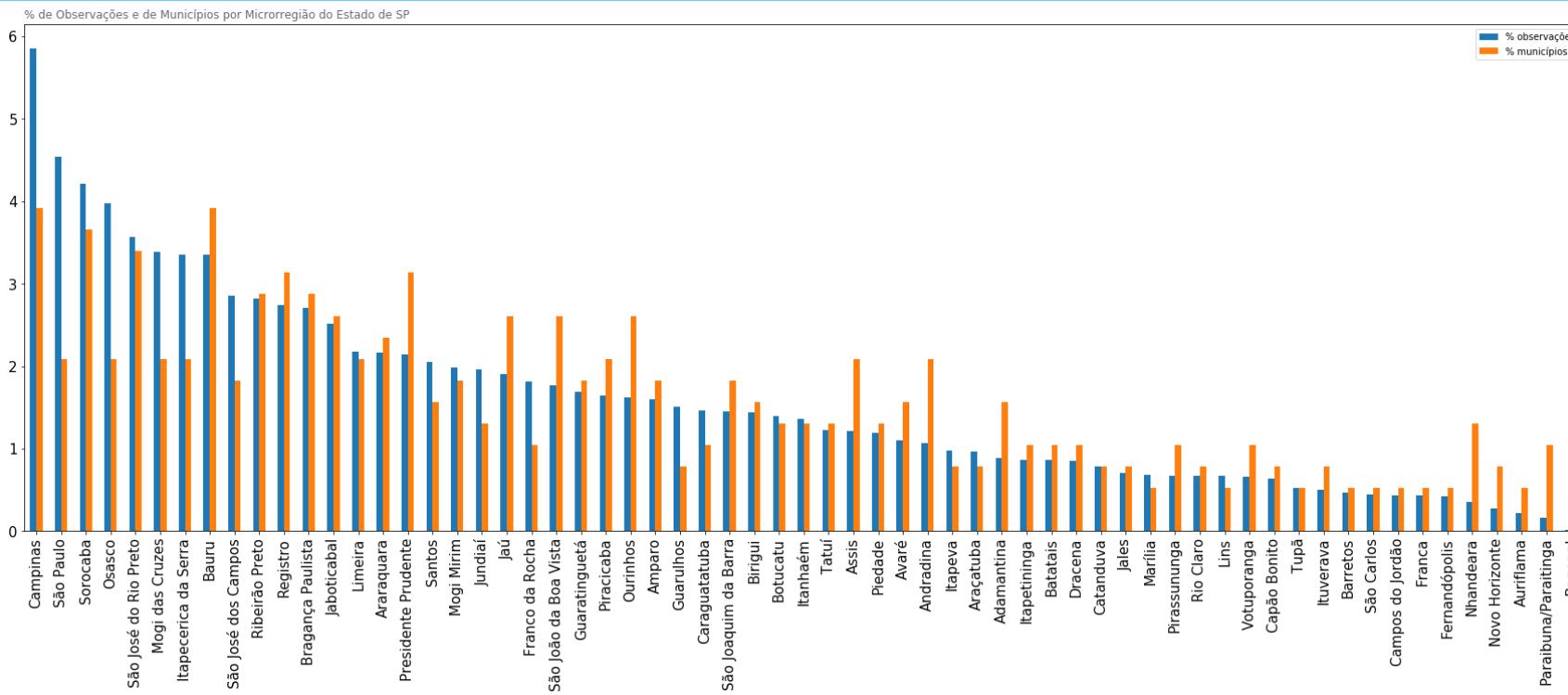


Detalhes das análises

26

## 'Microrregião Geográfica', 'Nome\_Microrregião'

- Na análise por microrregião, se destacam negativamente as microrregiões de:
  - Campinas, com quase 50% mais observações em relação ao percentual de municípios;
  - São Paulo, com 117% mais observações em relação ao percentual de municípios;
  - Osasco, com 90% mais observações em relação ao percentual de municípios.
- Isso demonstra que a doença esteve mais presente e de forma menos controlada nessas microrregiões até a data de corte, lembrando que São Paulo e Osasco pertencem à mesorregião Metropolitana de São Paulo



# 4. Análise Exploratória de Dados

RAIO-X DA BASE | ANÁLISE UNIVARIADA

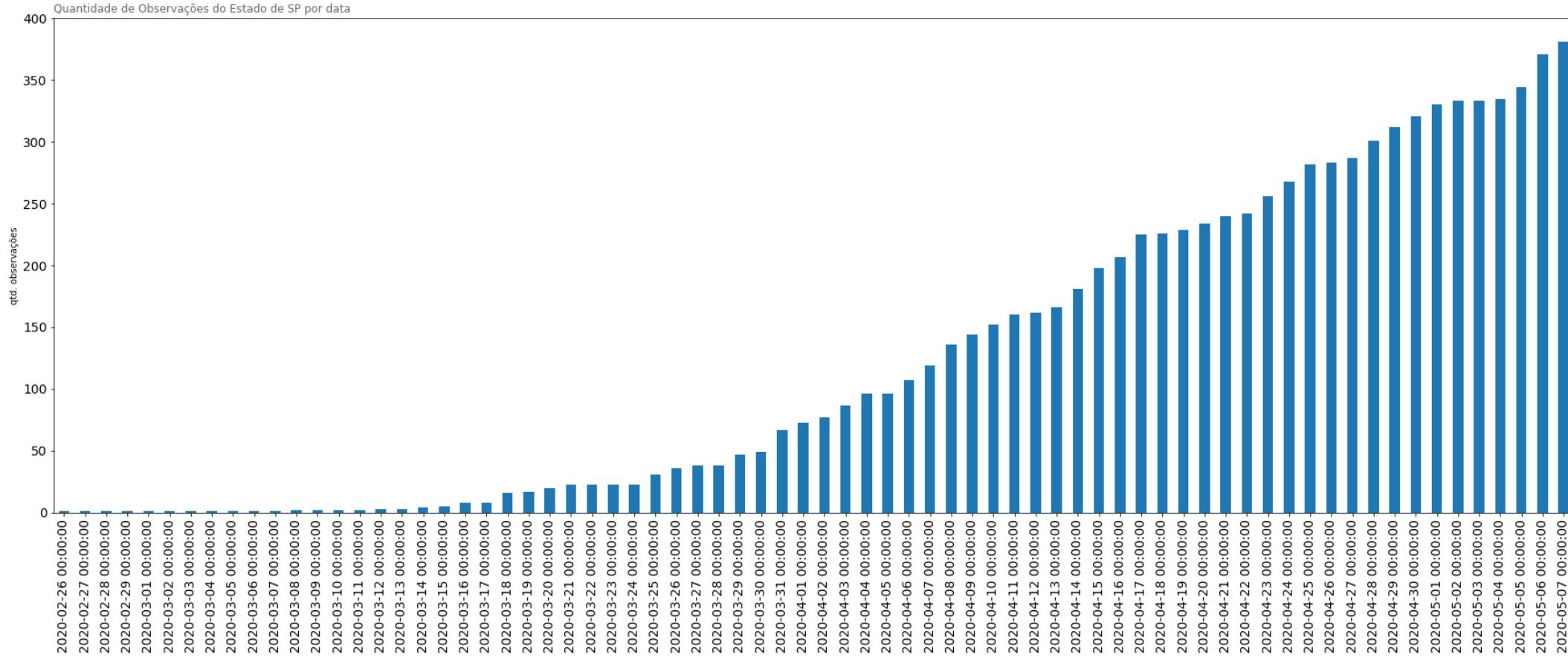


Detalhes das análises

27

'data'

- Ao analisarmos a evolução da quantidade de observações por data, podemos ver claramente a curva ascendente de novos municípios sendo atingidos pela pandemia.



# 4. Análise Exploratória de Dados

RAIO-X DA BASE | ANÁLISE UNIVARIADA

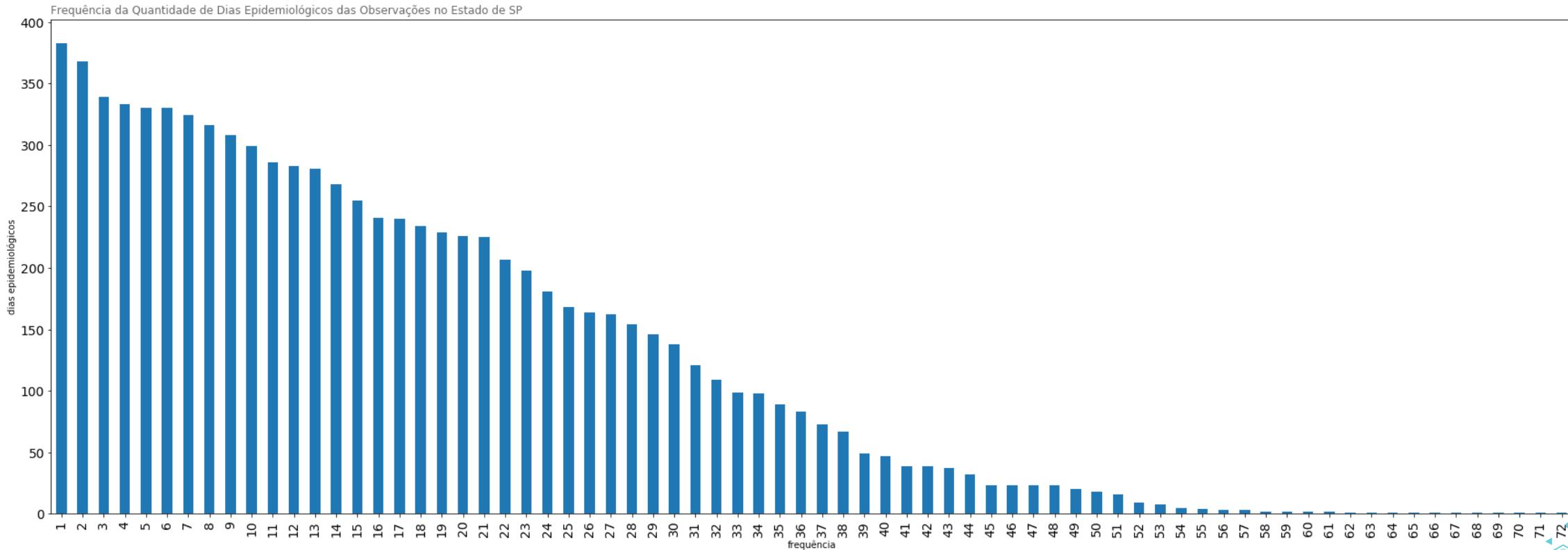


Detalhes das análises

28

## 'dias\_epidemiológicos'

- Ao analisarmos a frequência da quantidade de dias epidemiológicos, é natural vermos uma curva descendente conforme a quantidade de dias epidemiológicos aumenta, já que é da natureza da epidemia começar em uma cidade e, aos poucos, ir infectando outras.



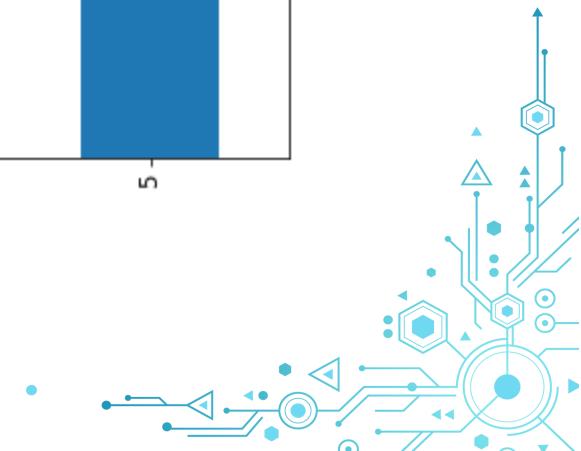
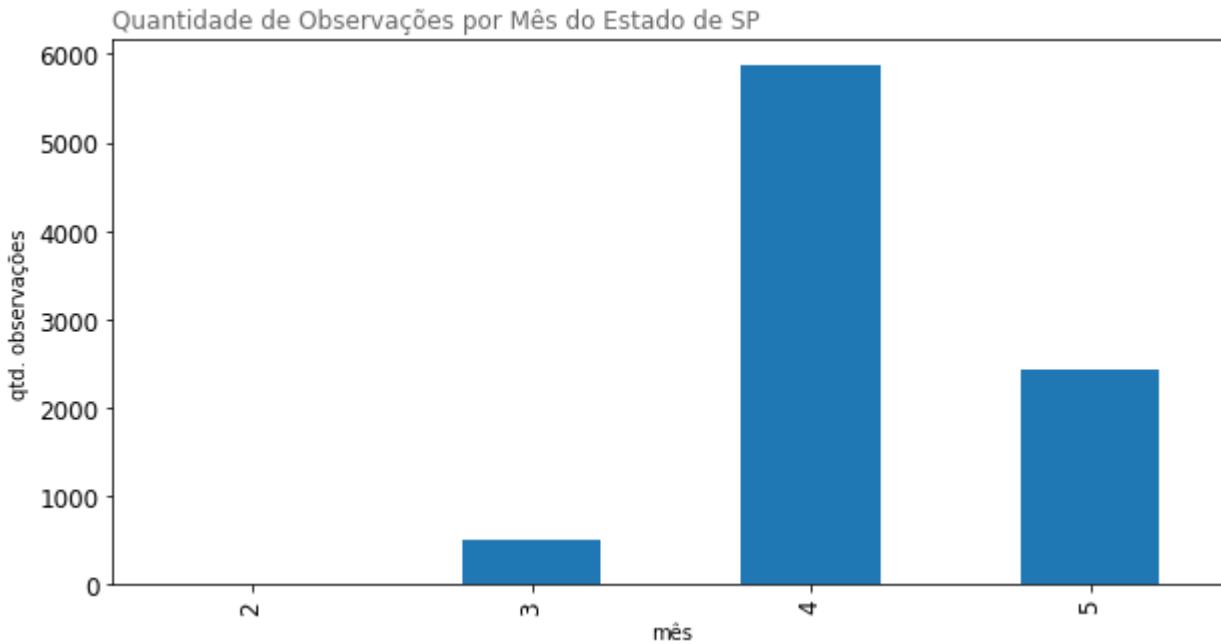
## 4. Análise Exploratória de Dados

RAIO-X DA BASE | ANÁLISE UNIVARIADA

### 'mês'

- Ao olhar o recorte do mês, é impressionante o salto de observações de março para abril, os únicos meses completos do dataset. E tão impressionante quanto é o mês de maio, que com apenas 7 dias já responde por 30% das observações.

mês	qtd.	%
2	4	0.045480
3	497	5.650938
4	5867	66.708357
5	2427	27.595225



# 4. Análise Exploratória de Dados

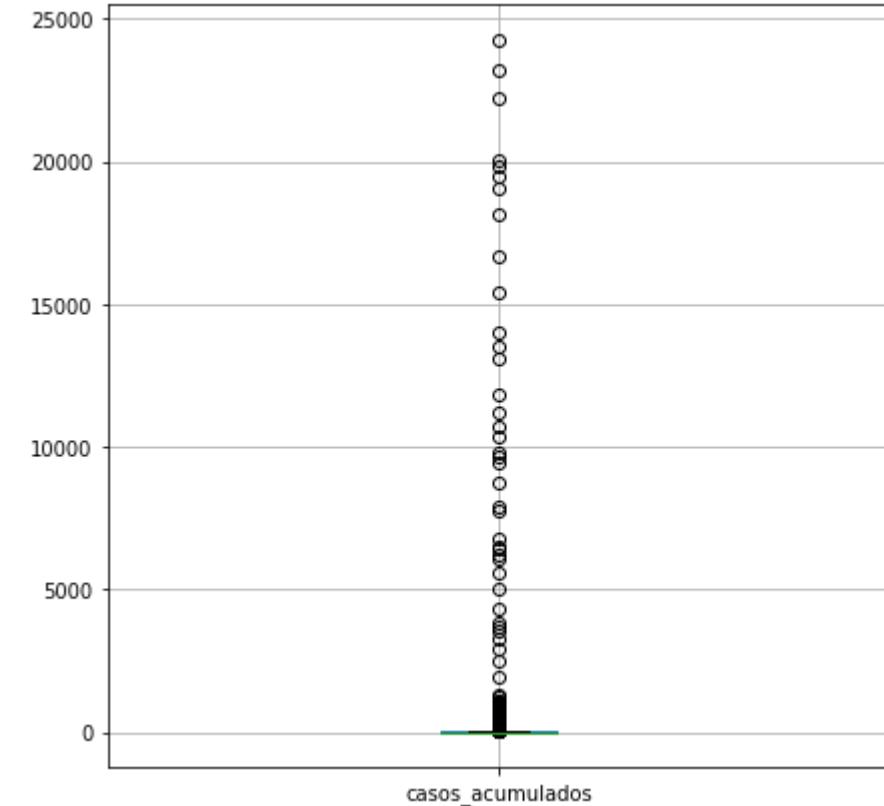
RAIO-X DA BASE | ANÁLISE UNIVARIADA

30

## 'casos acumulados'

- 50% das observações são de até 3 casos acumulados, o que faz sentido, pois os primeiros dias de transmissão num município costumam ser mais controlados, com baixa taxa de transmissão.
- A moda 1 também indica o início tímido da pandemia em cada município, que passam dias com um único caso acumulado até a transmissão se descontrolar e o número de casos explodir.
- O 3º quartil é de observações com até 13 casos acumulados.
- A partir daí, os casos acumulados costumam subir em progressão geométrica, donde que não consideramos outliers os pontos acima do 3º quartil, incluindo o ponto máximo, de 24273 casos acumulados.
- Não há casos acumulados negativos graças aos tratamentos realizados na base de dados, com investigação dos boletins diários divulgados pela Secretaria de Saúde estadual.
- Os altos valores de amplitude (24272), variância (669219) e coeficiente de variação (1150%) demonstram como os valores de casos diários estão espalhados e distantes da média (71).

	casos acumulados
Contagem	8795
Média	71.080500
Desvio Padrão	818.058554
Mínimo	1
25%	1
50%	3
75%	13
Máximo	24273
Moda	1
Mediana	3
Amplitude	24272
Variância	669219.797249
Coeficiente de Variação	1150.890259



# 4. Análise Exploratória de Dados

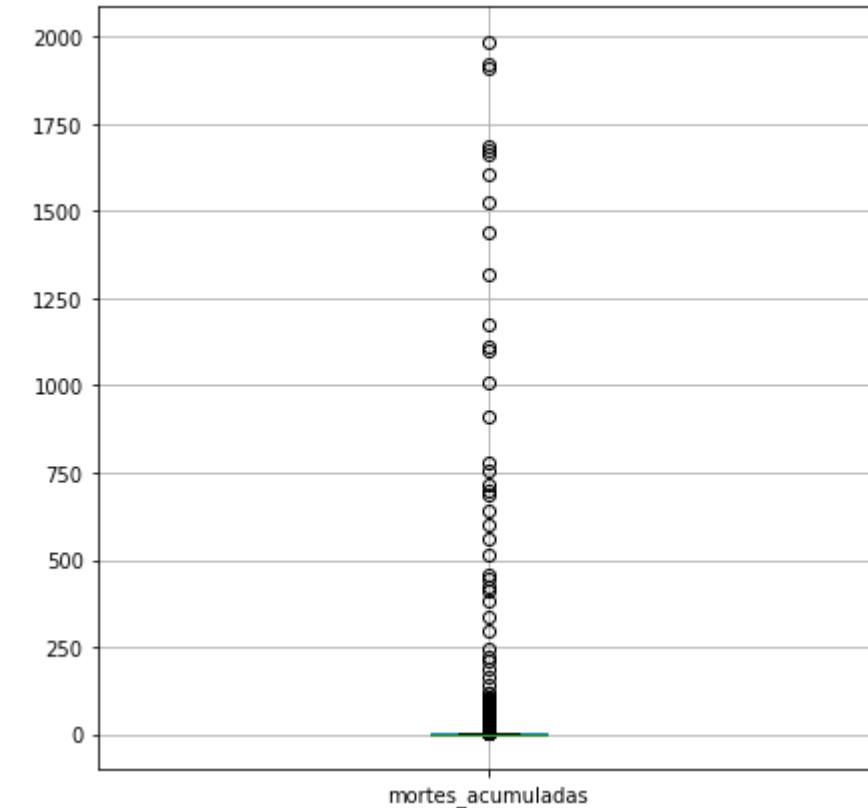
RAIO-X DA BASE | ANÁLISE UNIVARIADA

31

## 'mortes acumuladas'

- 50% das observações são de 0 mortes acumuladas, o que faz sentido, pois estima-se que a letalidade da COVID-19 gira em torno de 0.5% dos casos e as mortes só costumam aumentar quando há colapso do sistema de saúde.
- A moda e a mediana apontarem 0 também confirma o observado no boxplot.
- O 3º quartil aponta 1 morte acumulada.
- A partir daí, as mortes acumulados costumam subir em progressão geométrica, donde que não consideramos outliers os pontos acima do 3º quartil, incluindo o ponto máximo, de 1986 mortes acumuladas.
- Não há mortes acumuladas negativas graças aos tratamentos realizados na base de dados, com investigação dos boletins diários divulgados pela Secretaria de Saúde estadual.
- Os altos valores de amplitude (1986), variância (4495) e coeficiente de variação (1211%) demonstram como os valores de mortes acumuladas estão espalhados e distantes da média (5.5).

mortes acumuladas	
Contagem	8795
Média	5.533485
Desvio Padrão	67.051777
Mínimo	-9
25%	0
50%	0
75%	1
Máximo	1986
Moda	0
Mediana	0
Amplitude	1986
Variância	4495.940742
Coeficiente de Variação	1211.745896



## 4. Análise Exploratória de Dados

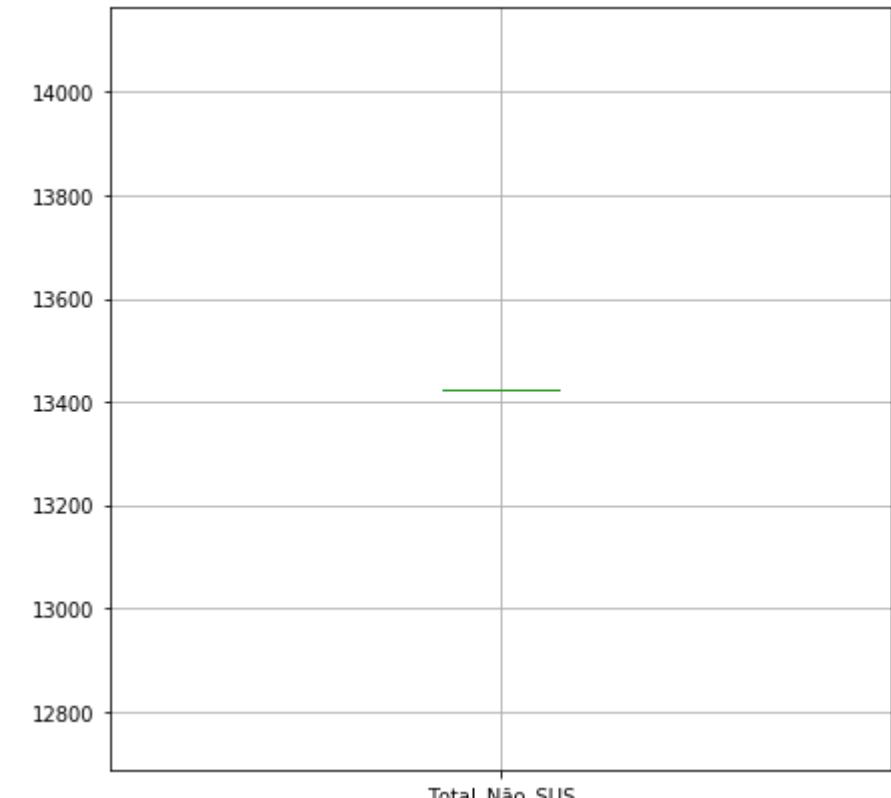
RAIO-X DA BASE | ANÁLISE UNIVARIADA

32

### 'Total de Leitos Não SUS' – mês 2

- Obs.: não faz sentido fazer análise univariada de nº de leitos hospitalares, pois eles são únicos para cada cidade e para cada mês. Portanto, vamos analisar o nº de leitos por mês.
- No mês 2, só havia observações da cidade de São Paulo, que apresentava 13425 leitos particulares.

	Total de Leitos Não SUS
Contagem	1
Média	13425
Desvio Padrão	-
Mínimo	13425
25%	13425
50%	13425
75%	13425
Máximo	13425
Soma	13425
Moda	13425
Mediana	13425
Amplitude	0
Variância	-
Coeficiente de Variação	-



# 4. Análise Exploratória de Dados

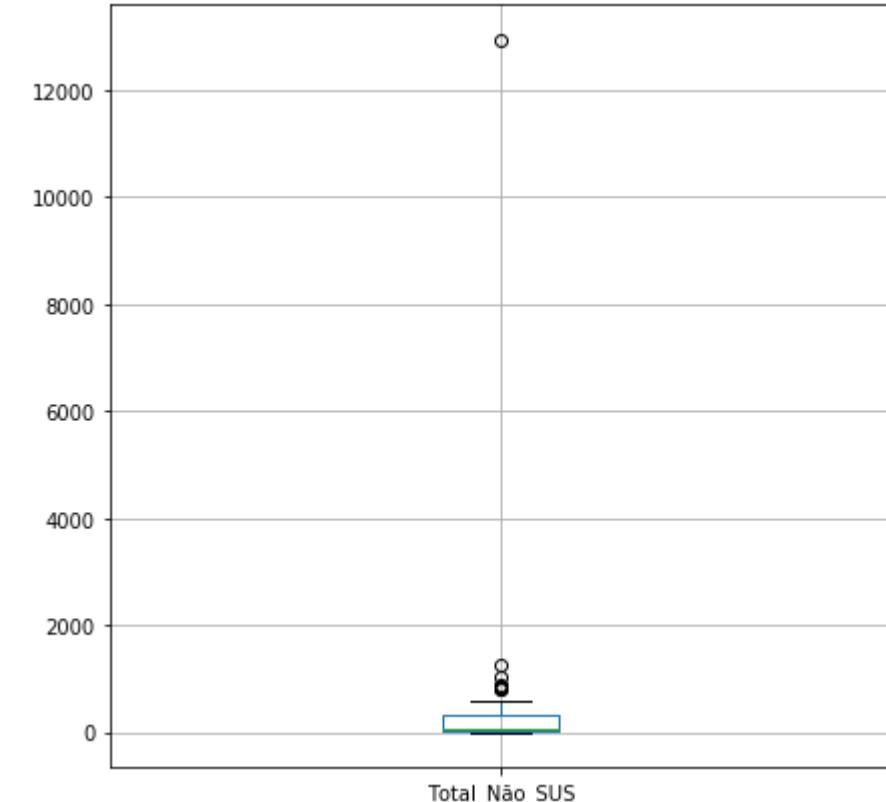
RAIO-X DA BASE | ANÁLISE UNIVARIADA

33

## 'Total de Leitos Não SUS' – mês 3

- No mês 3, havia 67 municípios do estado de São Paulo com ao menos 1 caso de COVID-19. Destes, a situação de leitos particulares foi a seguinte:
  - O 1º quartil dos 67 municípios possuía até 17 leitos particulares.
  - O 2º quartil dos 67 municípios possuía até 78 leitos particulares.
  - O 3º quartil dos 67 municípios possuía até 314 leitos particulares.
  - A cidade com maior oferta de leitos particulares foi São Paulo, com 12945 (o outlier no boxplot, que representa 89% da soma de todos os outros 66 municípios), o que denota queda em relação ao mês anterior.
  - Os 67 municípios paulistas com ao menos 1 caso de COVID-19 no mês 3 possuíam ao todo 27430 leitos particulares.
  - Os altos valores de amplitude (12945), variância (2504376) e coeficiente de variação (386%) demonstram como o nº de leitos particulares nos municípios está espalhado e distante da média (409).

Total de Leitos Não SUS	
Contagem	67
Média	409.40299
Desvio Padrão	1582.52233
Mínimo	0
25%	17
50%	78
75%	314.5
Máximo	12945
Soma	27430
Moda	0
Mediana	78
Amplitude	12945
Variância	2504376.9109
Coeficiente de Variação	386.54391



# 4. Análise Exploratória de Dados

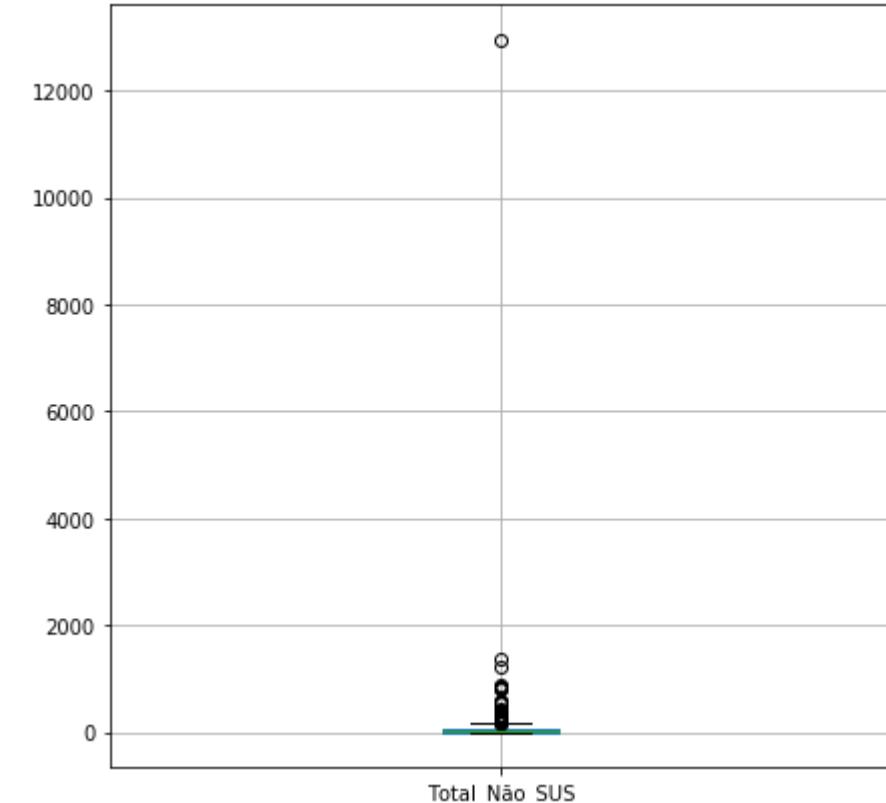
RAIO-X DA BASE | ANÁLISE UNIVARIADA

34

## 'Total de Leitos Não SUS' – mês 4

- No mês 4, havia 321 municípios do estado de São Paulo com ao menos 1 caso de COVID-19. Destes, a situação de leitos particulares foi a seguinte:
  - O 1º quartil dos 321 municípios não possuía leitos particulares.
  - O 2º quartil dos 321 municípios possuía até 14 leitos particulares.
  - O 3º quartil dos 321 municípios possuía até 66 leitos particulares.
  - A cidade com maior oferta de leitos particulares foi São Paulo, com 12955 (o outlier no boxplot, que representa 54% da soma de todos os outros 320 municípios), o que denota aumento em relação ao mês anterior.
  - Os 321 municípios paulistas com ao menos 1 caso de COVID-19 no mês 4 possuíam ao todo 36834 leitos particulares.
  - Os altos valores de amplitude (12955), variância (545879) e coeficiente de variação (643%) demonstram como o nº de leitos particulares nos municípios está espalhado e distante da média (114).

	Total de Leitos Não SUS
Contagem	321
Média	114.74766
Desvio Padrão	738.83687
Mínimo	0
25%	0
50%	14
75%	66
Máximo	12955
Soma	36834
Moda	0
Mediana	14
Amplitude	12955
Variância	545879.92050
Coeficiente de Variação	643.87966



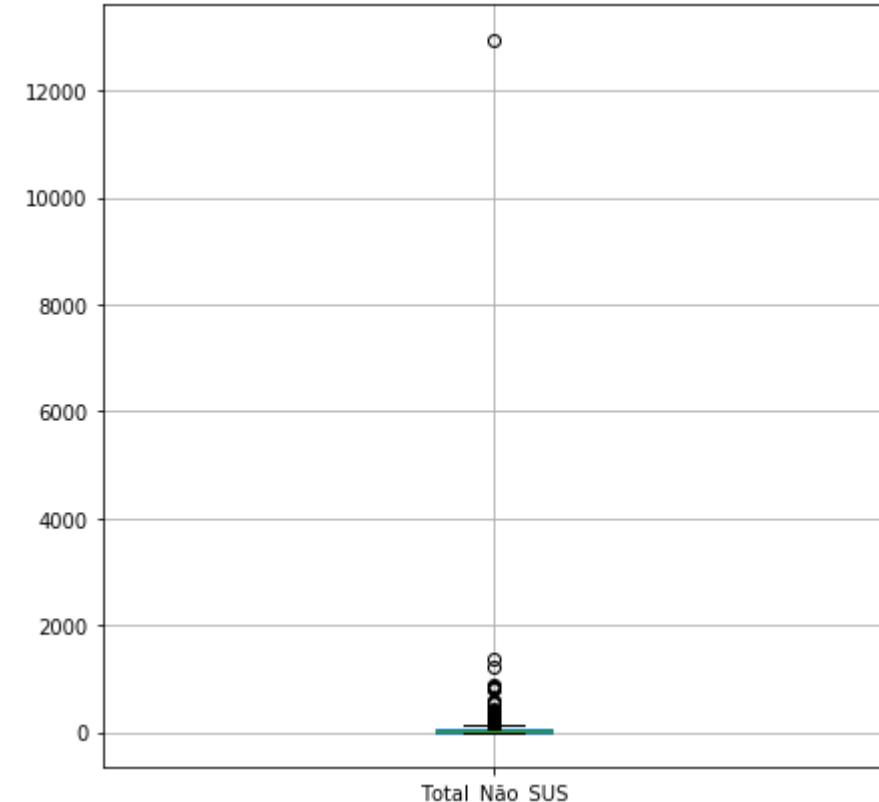
## 4. Análise Exploratória de Dados

RAIO-X DA BASE | ANÁLISE UNIVARIADA

### 'Total de Leitos Não SUS' – mês 5

- No mês 5, havia 383 municípios do estado de São Paulo com ao menos 1 caso de COVID-19. Destes, a situação de leitos particulares foi a seguinte:
- O 1º quartil dos 383 municípios não possuía leitos particulares.
- O 2º quartil dos 383 municípios possuía até 10 leitos particulares.
- O 3º quartil dos 383 municípios possuía até 50 leitos particulares.
- A cidade com maior oferta de leitos particulares foi São Paulo, com 12955 (o outlier no boxplot, que representa 53% da soma de todos os outros 382 municípios), ou seja, São Paulo manteve o mesmo nº de leitos particulares em relação ao mês anterior.
- Os 383 municípios paulistas com ao menos 1 caso de COVID-19 no mês 5 possuíam ao todo 37159 leitos particulares.
- Os altos valores de amplitude (12955), variância (458928) e coeficiente de variação (698%) demonstram como o nº de leitos particulares nos municípios está espalhado e distante da média (97).

	Total de Leitos Não SUS
Contagem	383
Média	97.02089
Desvio Padrão	677.44271
Mínimo	0
25%	0
50%	10
75%	50
Máximo	12955
Soma	37159
Moda	0
Mediana	10
Amplitude	12955
Variância	458928.62260
Coeficiente de Variação	698.24419



Detalhes das análises

## 4. Análise Exploratória de Dados

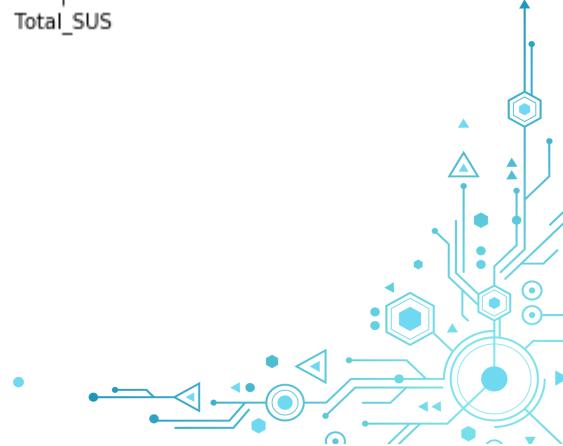
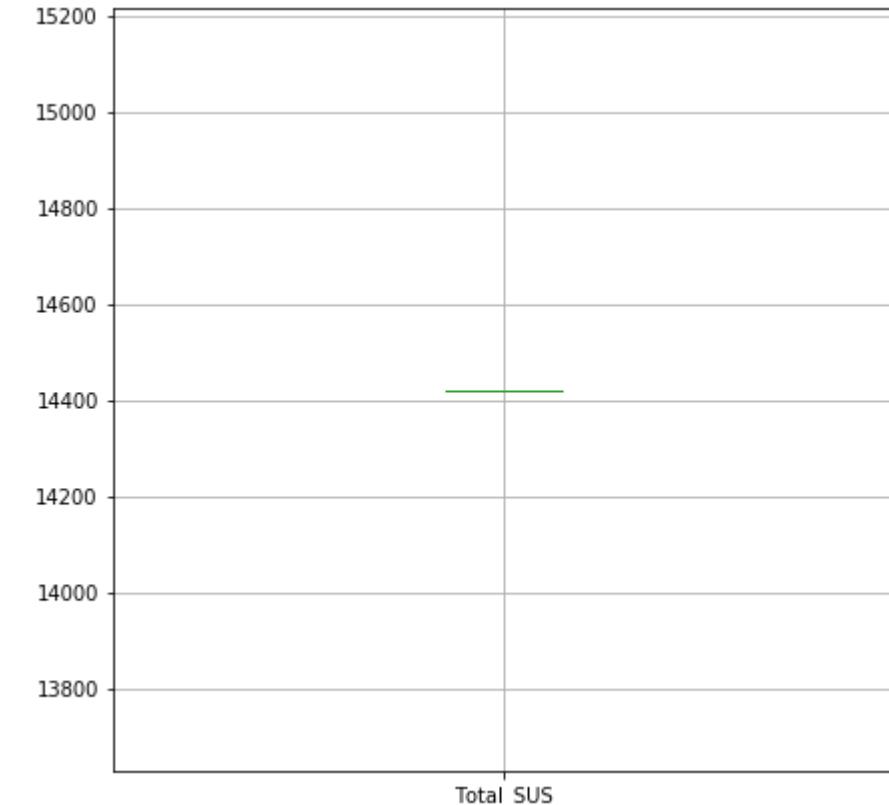
RAIO-X DA BASE | ANÁLISE UNIVARIADA

36

### 'Total de Leitos SUS' – mês 2

- Obs.: não faz sentido fazer análise univariada de nº de leitos hospitalares, pois eles são únicos para cada cidade e para cada mês. Portanto, vamos analisar o nº de leitos por mês.
- No mês 2, só havia observações da cidade de São Paulo, que apresentava 14422 leitos do SUS.

	Total de Leitos SUS
Contagem	1
Média	14422
Desvio Padrão	-
Mínimo	14422
25%	14422
50%	14422
75%	14422
Máximo	14422
Soma	14422
Moda	14422
Mediana	14422
Amplitude	0
Variância	-
Coeficiente de Variação	-



# 4. Análise Exploratória de Dados

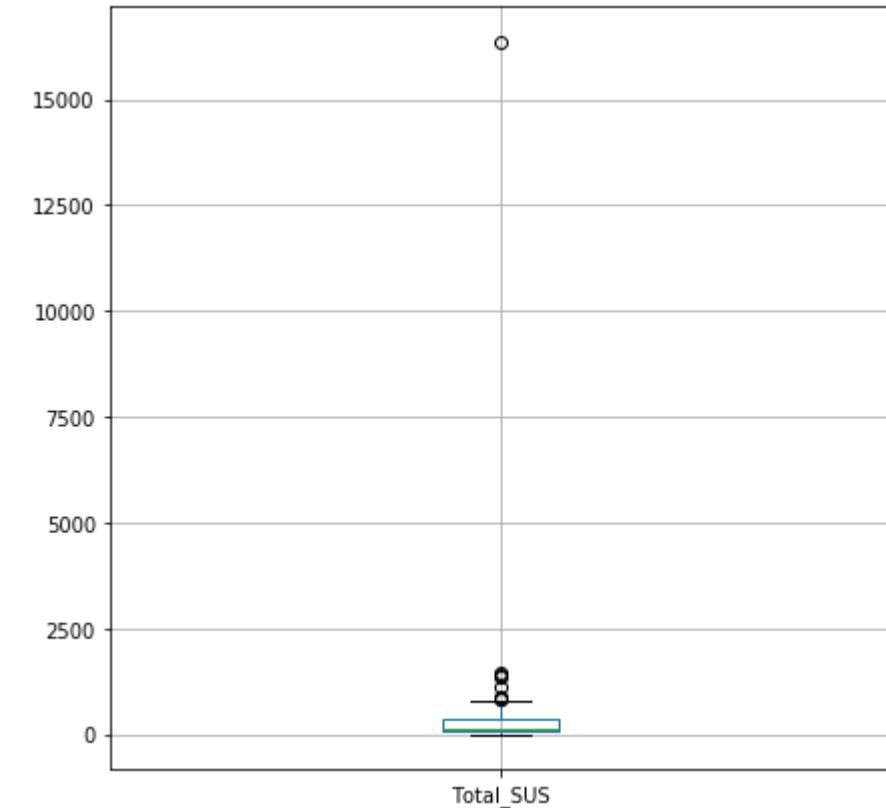
RAIO-X DA BASE | ANÁLISE UNIVARIADA

37

## 'Total de Leitos SUS' – mês 3

- No mês 3, havia 67 municípios do estado de São Paulo com ao menos 1 caso de COVID-19. Destes, a situação de leitos do SUS foi a seguinte:
  - O 1º quartil dos 67 municípios possuía até 55 leitos do SUS.
  - O 2º quartil dos 67 municípios possuía até 115 leitos do SUS.
  - O 3º quartil dos 67 municípios possuía até 342 leitos do SUS.
  - A cidade com maior oferta de leitos do SUS foi São Paulo, com 16367 (o outlier no boxplot, que representa 90% da soma de todos os outros 66 municípios), o que denota aumento em relação ao mês anterior.
  - Os 67 municípios paulistas com ao menos 1 caso de COVID-19 no mês 3 possuíam ao todo 34489 leitos do SUS.
  - Os altos valores de amplitude (16367), variância (3989245) e coeficiente de variação (388%) demonstram como o nº de leitos do SUS nos municípios está espalhado e distante da média (514).

	Total de Leitos SUS
Contagem	67
Média	514.76119
Desvio Padrão	1997.30952
Mínimo	0
25%	55
50%	115
75%	342.5
Máximo	16367
Soma	34489
Moda	0, 78, 141 e 189
Mediana	115
Amplitude	16367
Variância	3989245.3360
Coeficiente de Variação	388.00701



# 4. Análise Exploratória de Dados

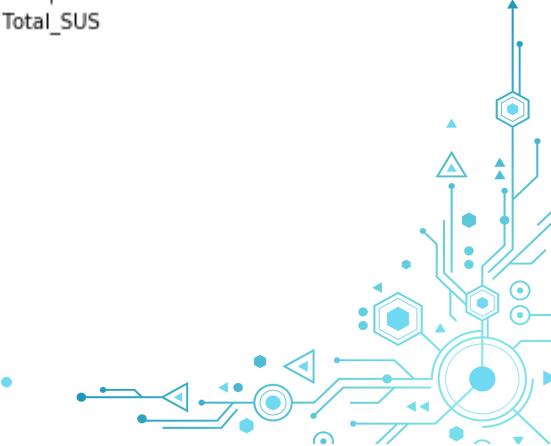
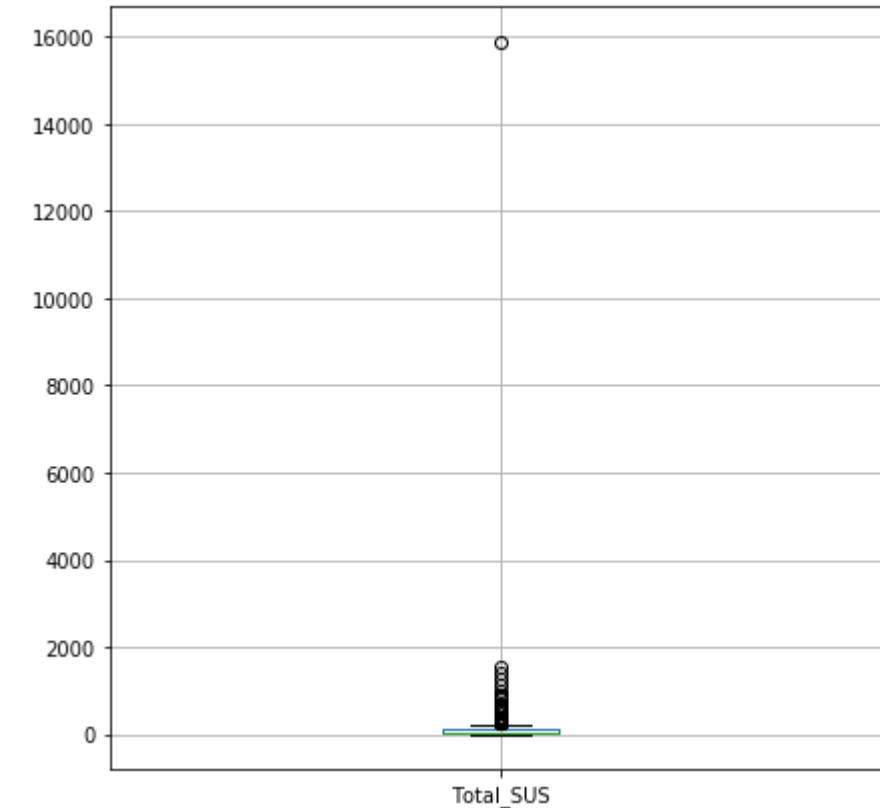
RAIO-X DA BASE | ANÁLISE UNIVARIADA

38

## 'Total de Leitos SUS' – mês 4

- No mês 4, havia 321 municípios do estado de São Paulo com ao menos 1 caso de COVID-19. Destes, a situação de leitos do SUS foi a seguinte:
  - O 1º quartil dos 321 municípios possuía até 9 leitos do SUS.
  - O 2º quartil dos 321 municípios possuía até 43 leitos do SUS.
  - O 3º quartil dos 321 municípios possuía até 101 leitos do SUS.
  - A cidade com maior oferta de leitos do SUS foi São Paulo, com 15893 (o outlier no boxplot, que representa 42% da soma de todos os outros 320 municípios), o que denota queda em relação ao mês anterior.
  - Os 321 municípios paulistas com ao menos 1 caso de COVID-19 no mês 4 possuíam ao todo 53583 leitos do SUS.
  - Os altos valores de amplitude (15893), variância (822446) e coeficiente de variação (543%) demonstram como o nº de leitos do SUS nos municípios está espalhado e distante da média (166).

Total de Leitos SUS	
Contagem	321
Média	166.92523
Desvio Padrão	906.88850
Mínimo	0
25%	9
50%	43
75%	101
Máximo	15893
Soma	53583
Moda	0
Mediana	43
Amplitude	15893
Variância	822446.74439
Coeficiente de Variação	543.29024



## 4. Análise Exploratória de Dados

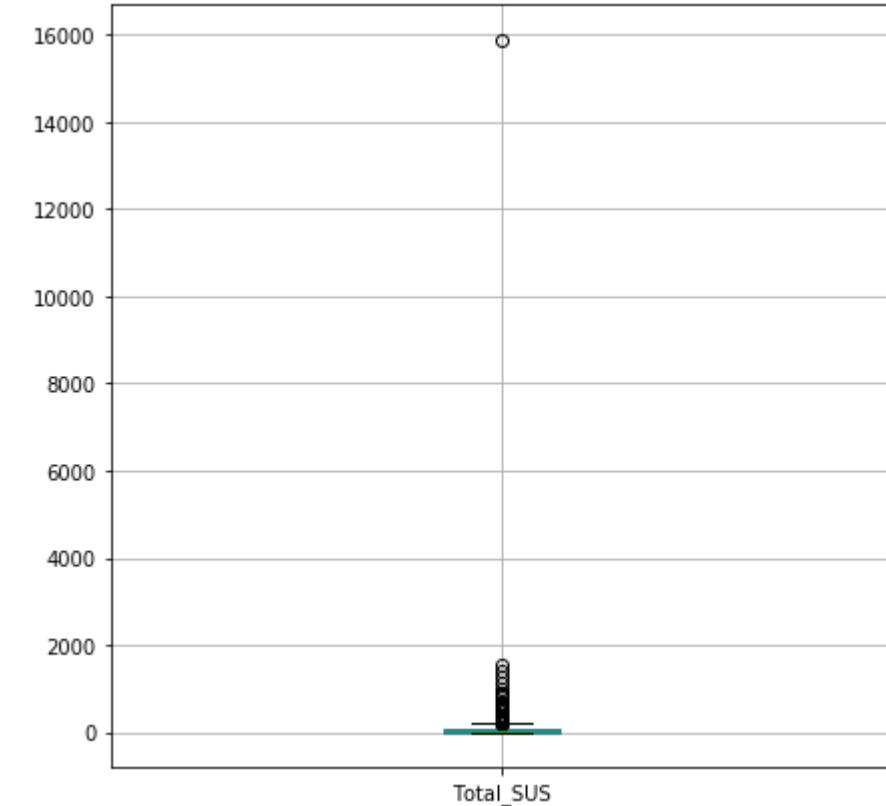
RAIO-X DA BASE | ANÁLISE UNIVARIADA

39

### 'Total de Leitos SUS' – mês 5

- No mês 5, havia 383 municípios do estado de São Paulo com ao menos 1 caso de COVID-19. Destes, a situação de leitos do SUS foi a seguinte:
  - O 1º quartil dos 383 municípios não possuía leitos do SUS.
  - O 2º quartil dos 383 municípios possuía até 32 leitos do SUS.
  - O 3º quartil dos 383 municípios possuía até 84 leitos do SUS.
  - A cidade com maior oferta de leitos do SUS foi São Paulo, com 15893 (o outlier no boxplot, que representa 41% da soma de todos os outros 382 municípios), ou seja, São Paulo manteve o mesmo nº de leitos do SUS em relação ao mês anterior.
  - Os 383 municípios paulistas com ao menos 1 caso de COVID-19 no mês 5 possuíam ao todo 54419 leitos do SUS.
  - Os altos valores de amplitude (15893), variância (692204) e coeficiente de variação (585%) demonstram como o nº de leitos do SUS nos municípios está espalhado e distante da média (142).

Total de Leitos SUS	
Contagem	383
Média	142.08616
Desvio Padrão	831.98842
Mínimo	0
25%	0
50%	32
75%	84
Máximo	15893
Soma	54419
Moda	0
Mediana	32
Amplitude	15893
Variância	692204.72292
Coeficiente de Variação	585.55204



Detalhes das análises

# 4. Análise Exploratória de Dados

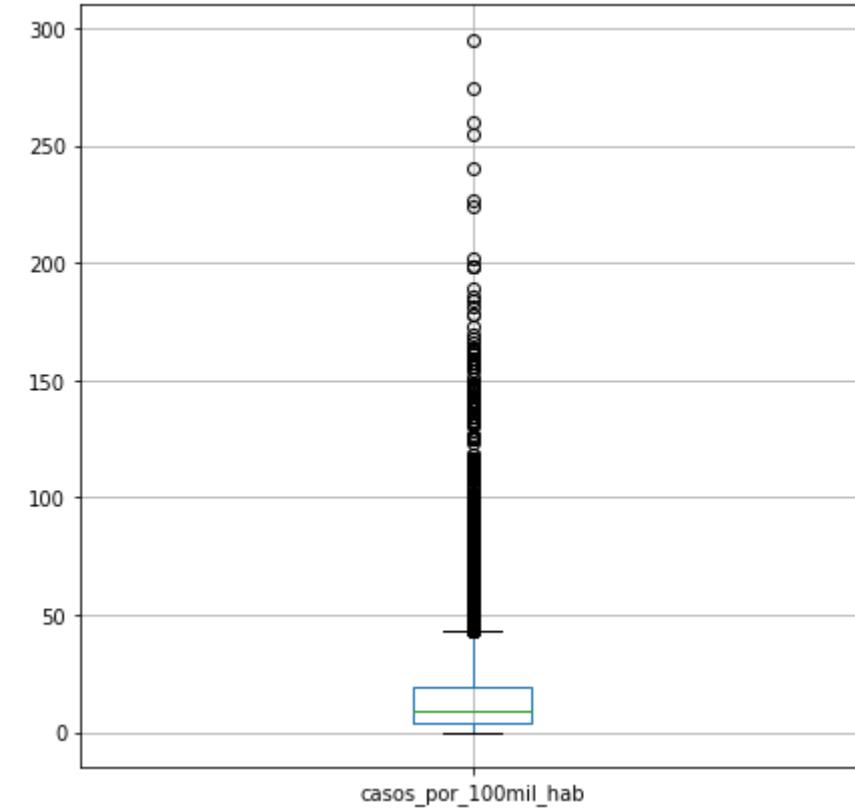
RAIO-X DA BASE | ANÁLISE UNIVARIADA

40

## 'Casos por 100 mil habitantes'

- O 1º quartil é de observações com até 3.6 casos por 100 mil habitantes.
- O 2º quartil é de observações com até 8.8 casos por 100 mil habitantes.
- O 3º quartil é de observações com até 19.4 casos por 100 mil habitantes.
- Essa distribuição dos quartis com número relativamente baixo de casos por 100 mil habitantes demonstra que a pandemia ainda estava em seu estágio inicial até a data de corte (7/5).
- A moda 1.7 indica o início tímido da pandemia em cada município, que passam dias com um único caso acumulado até a transmissão se descontrolar e o número de casos explodir.
- Não consideramos outliers os pontos acima de 50 casos por 100 mil habitantes, já que a tendência de contaminação é a subida em progressão geométrica.
- Os altos valores de amplitude (295), variância (565) e coeficiente de variação (143%) demonstram como os valores de casos por 100 mil habitantes estão espalhados e distantes da média (16).

Casos por 100 mil habitantes	
Contagem	8795
Média	16.59065
Desvio Padrão	23.78735
Mínimo	0
25%	3.6
50%	8.8
75%	19.4
Máximo	295.2
Moda	1.7
Mediana	8.8
Amplitude	295.2
Variância	565.83779
Coeficiente de Variação	143.37798



# 4. Análise Exploratória de Dados

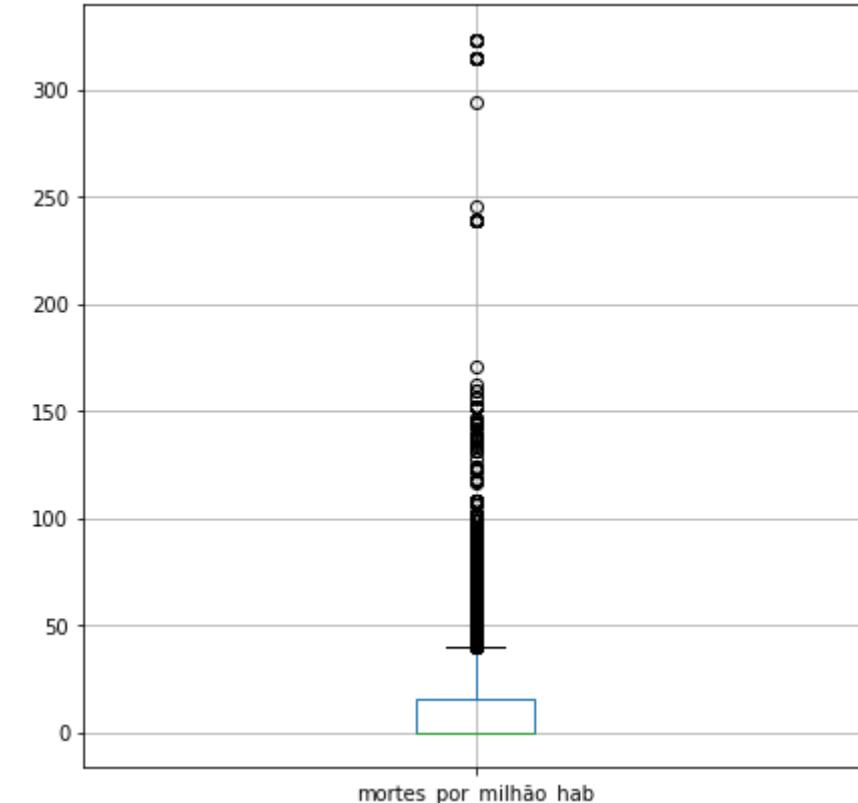
RAIO-X DA BASE | ANÁLISE UNIVARIADA

41

## 'Mortes por milhão de habitantes'

- Metade das observações não tem mortes, o que é explicado devido ao estágio inicial da pandemia até a data de corte (7/5): como nem todo caso de infecção pelo SARS-CoV-2 leva necessariamente à morte e como a COVID-19 é tratável, as mortes costumam acontecer quando há muitas pessoas infectadas e/ou quando há colapso do sistema de saúde.
- O 3º quartil é de observações com até 15.9 casos por milhão de habitantes.
- A moda 0 indica o início tímido da pandemia em cada município, que passam dias após o 1º caso até a 1ª morte.
- Não consideramos outliers os pontos acima de 50 mortes por milhão de habitantes, já que a tendência de mortes é a subida em progressão geométrica quando não há controle dos casos de infecção ou devido investimento no sistema de saúde.
- Os altos valores de amplitude (323), variância (977) e coeficiente de variação (226%) demonstram como os valores de mortes por milhão de habitantes estão espalhados e distantes da média (13).

Mortes por milhão de habitantes	
Contagem	8795
Média	13.82070
Desvio Padrão	31.27207
Mínimo	0
25%	0
50%	0
75%	15.9
Máximo	323.3
Moda	0
Mediana	0
Amplitude	323.3
Variância	977.94262
Coeficiente de Variação	226.26975



# 4. Análise Exploratória de Dados

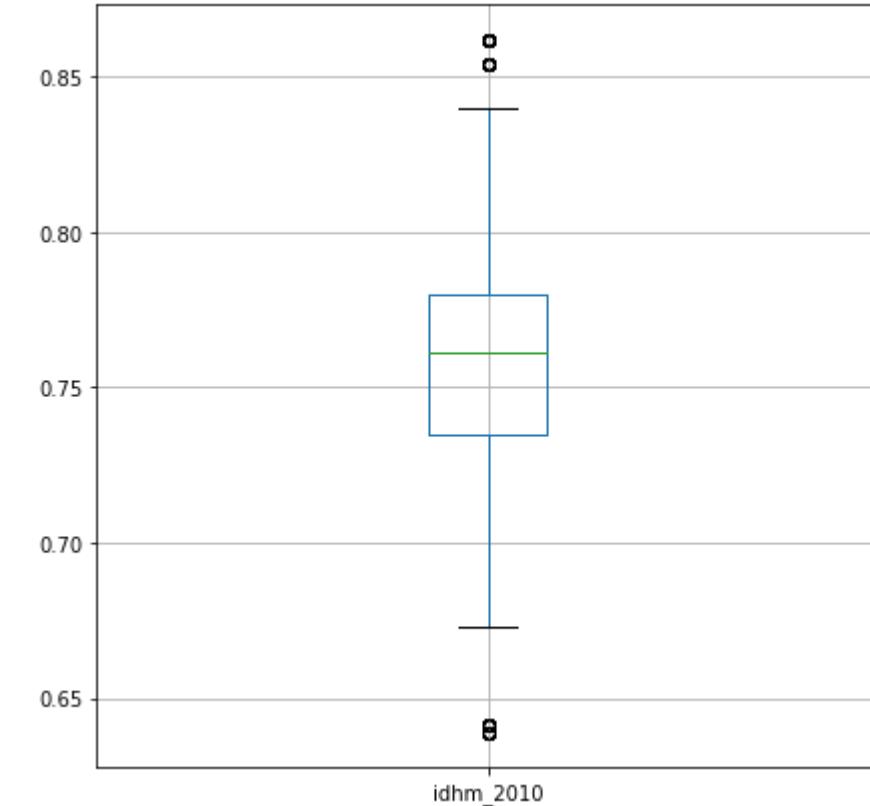
RAIO-X DA BASE | ANÁLISE UNIVARIADA

42

## 'IDHM 2010'

- O 1º quartil é de observações em municípios com IDH 2010 até 0.73500.
- O 2º quartil é de observações em municípios com IDH 2010 até 0.76100.
- O 3º quartil é de observações em municípios com IDH 2010 até 0.78000.
- O município de menor IDH 2010 no estado de São Paulo com casos de COVID-19 é Ribeirão Branco, cujo índice é 0.63900.
- O município de maior IDH 2010 no estado de São Paulo com casos de COVID-19 é São Caetano do Sul, cujo índice é 0.86200.
- A título de comparação, a capital de São Paulo tem IDH 2010 de 0.80500.
- Os baixos valores de amplitude (0.22300), variância (0.00115) e coeficiente de variação (4.46903%) demonstram como os valores de IDH 2010 dos municípios do Estado de São Paulo estão juntos e próximos da média (0.75819).
- Uma das hipóteses que este estudo pretende verificar é a importância do índice de desenvolvimento humano municipal na propagação dos casos ou das mortes por COVID-19.

IDHM 2010	
Contagem	8795
Média	0.75819
Desvio Padrão	0.03388
Mínimo	0.63900
25%	0.73500
50%	0.76100
75%	0.78000
Máximo	0.86200
Moda	0.74900
Mediana	0.76100
Amplitude	0.22300
Variância	0.00115
Coeficiente de Variação	4.46903



# 4. Análise Exploratória de Dados

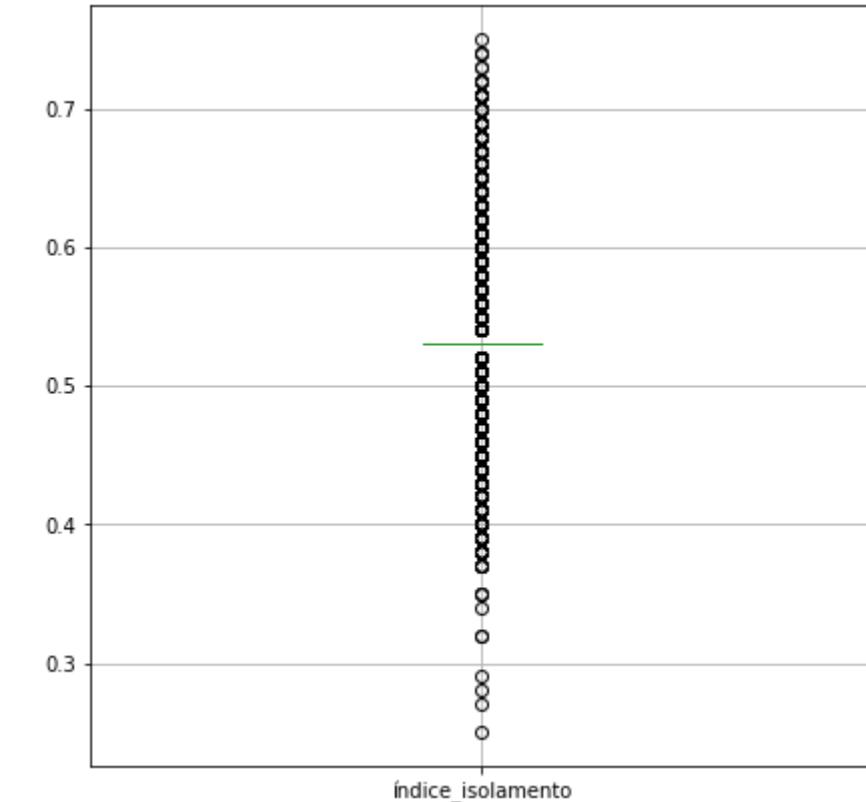
RAIO-X DA BASE | ANÁLISE UNIVARIADA

43

## 'Índice de Isolamento'

- Como pouco mais da metade das cidades paulistas não possuíam índice de isolamento, consideramos o valor da mediana do índice de isolamento no lugar dos valores faltantes. Por isso, 3/4 das observações têm índice de isolamento até 0.53.
- A cidade que registrou menor índice de isolamento foi São Paulo, em 12/03, com 0.25.
- A cidade que registrou maior índice de isolamento foi São Sebastião, em 04/04, com 0.75.
- Os baixos valores de amplitude (0.50000), variância (0.00183) e coeficiente de variação (8.05846%) demonstram como os valores do índice de isolamento estão juntos e próximos da média (0.53036).

Índice de Isolamento	
Contagem	8795
Média	0.53036
Desvio Padrão	0.04274
Mínimo	0.25000
25%	0.53000
50%	0.53000
75%	0.53000
Máximo	0.75000
Moda	0.53000
Mediana	0.53000
Amplitude	0.50000
Variância	0.00183
Coeficiente de Variação	8.05846



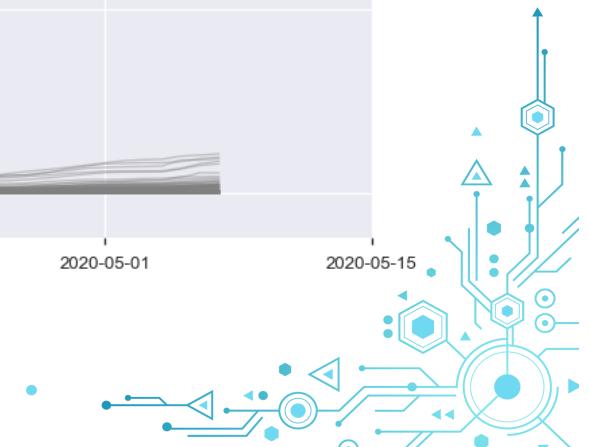
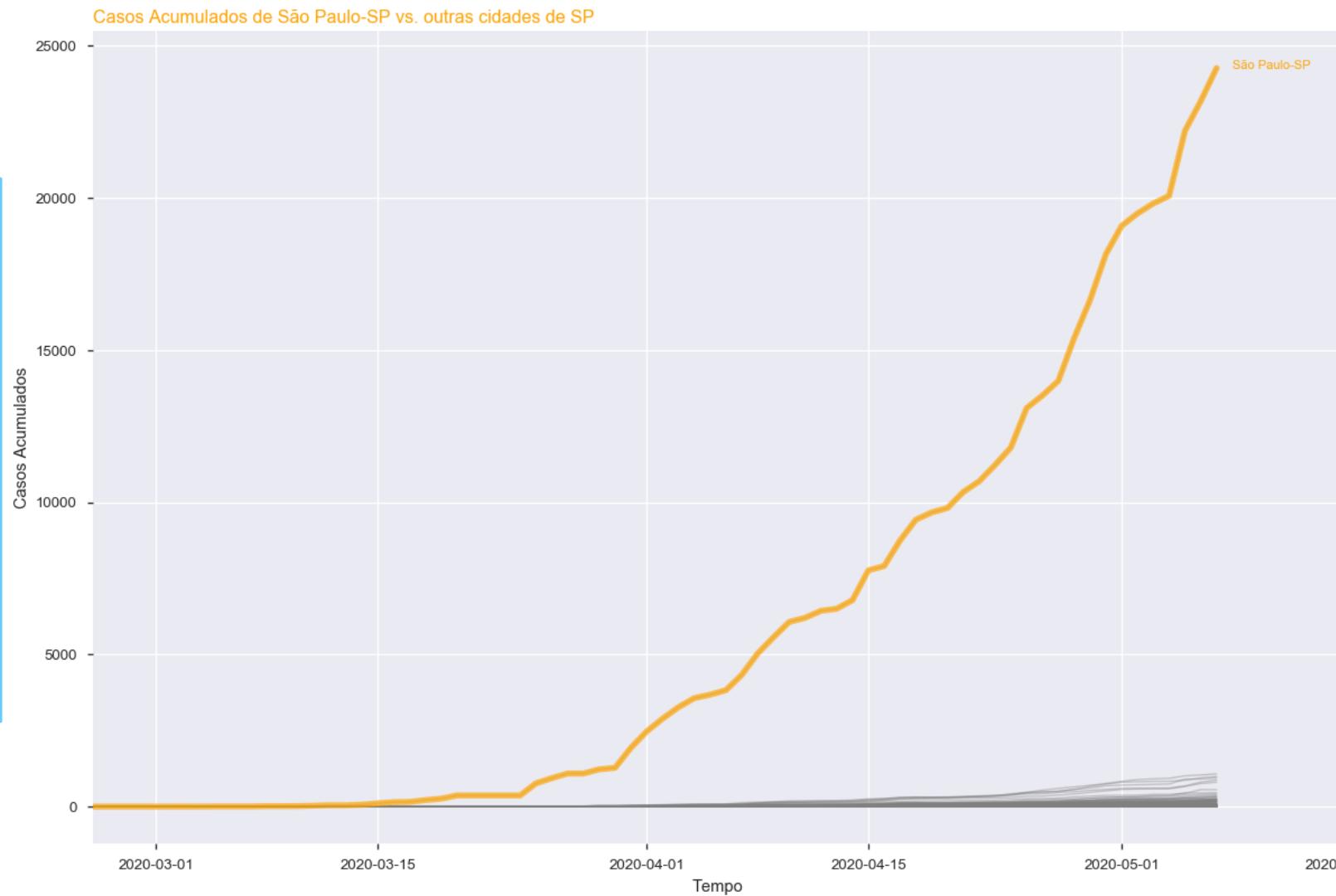
## 4. Análise Exploratória de Dados

RAIO-X DA BASE | ANÁLISE MULTIVARIADA

44

### Casos Acumulados por Cidade e por Data

- Aqui, vemos como a pandemia se inicia na capital e explode em nº de casos em relação aos outros municípios do estado.



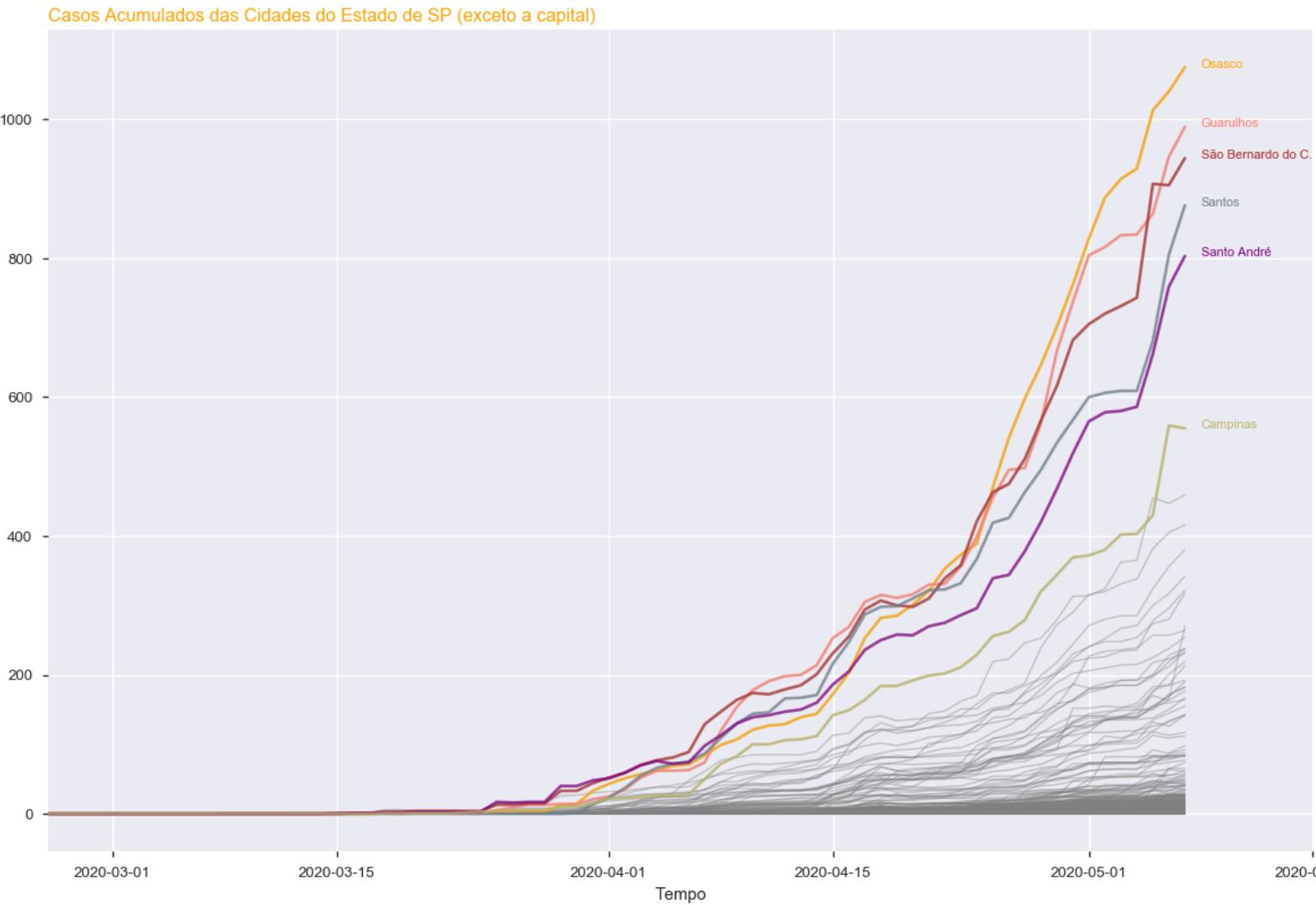
# 4. Análise Exploratória de Dados

RAIO-X DA BASE | ANÁLISE MULTIVARIADA

45

## Casos Acumulados por Cidade e por Data

- Aqui, vemos o aumento no nº de casos no interior de SP.



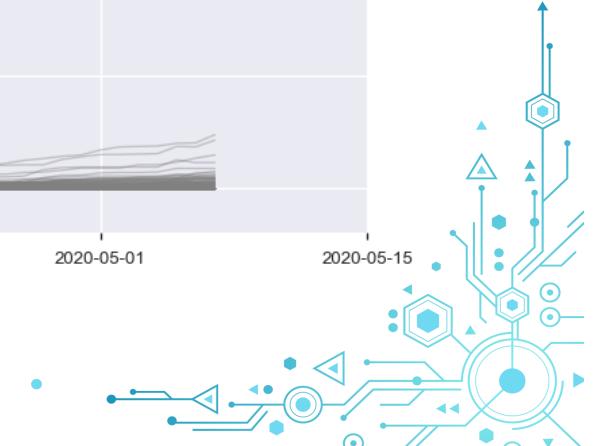
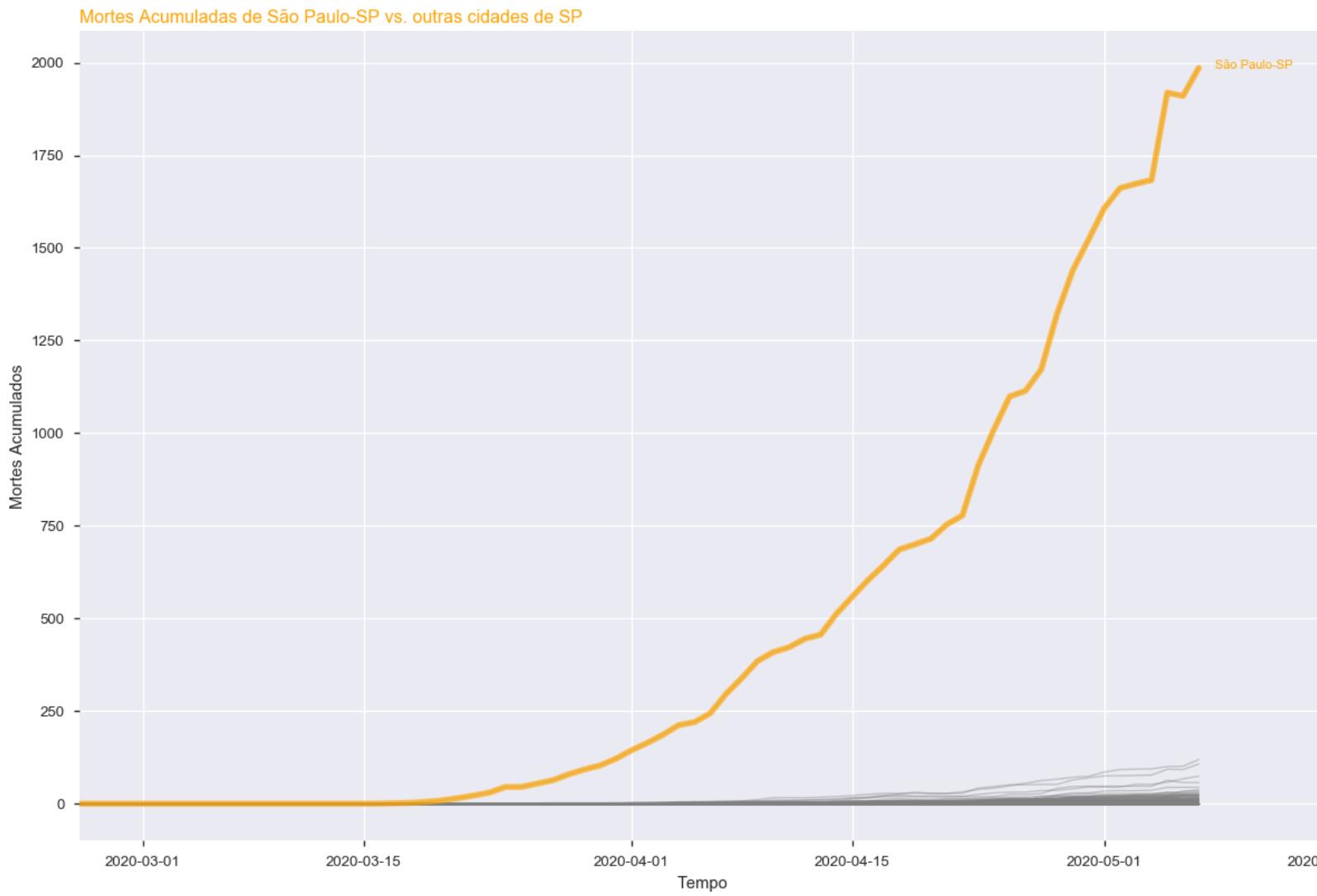
# 4. Análise Exploratória de Dados

RAIO-X DA BASE | ANÁLISE MULTIVARIADA

46

## Mortes Acumuladas por Cidade e por Data

- São Paulo acumulando quase 2000 mortes em 72 dias.



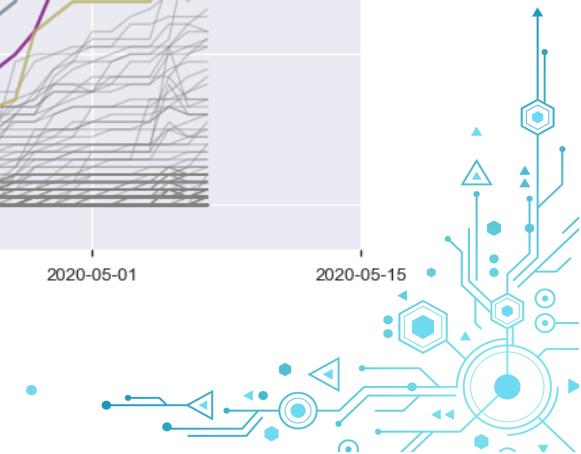
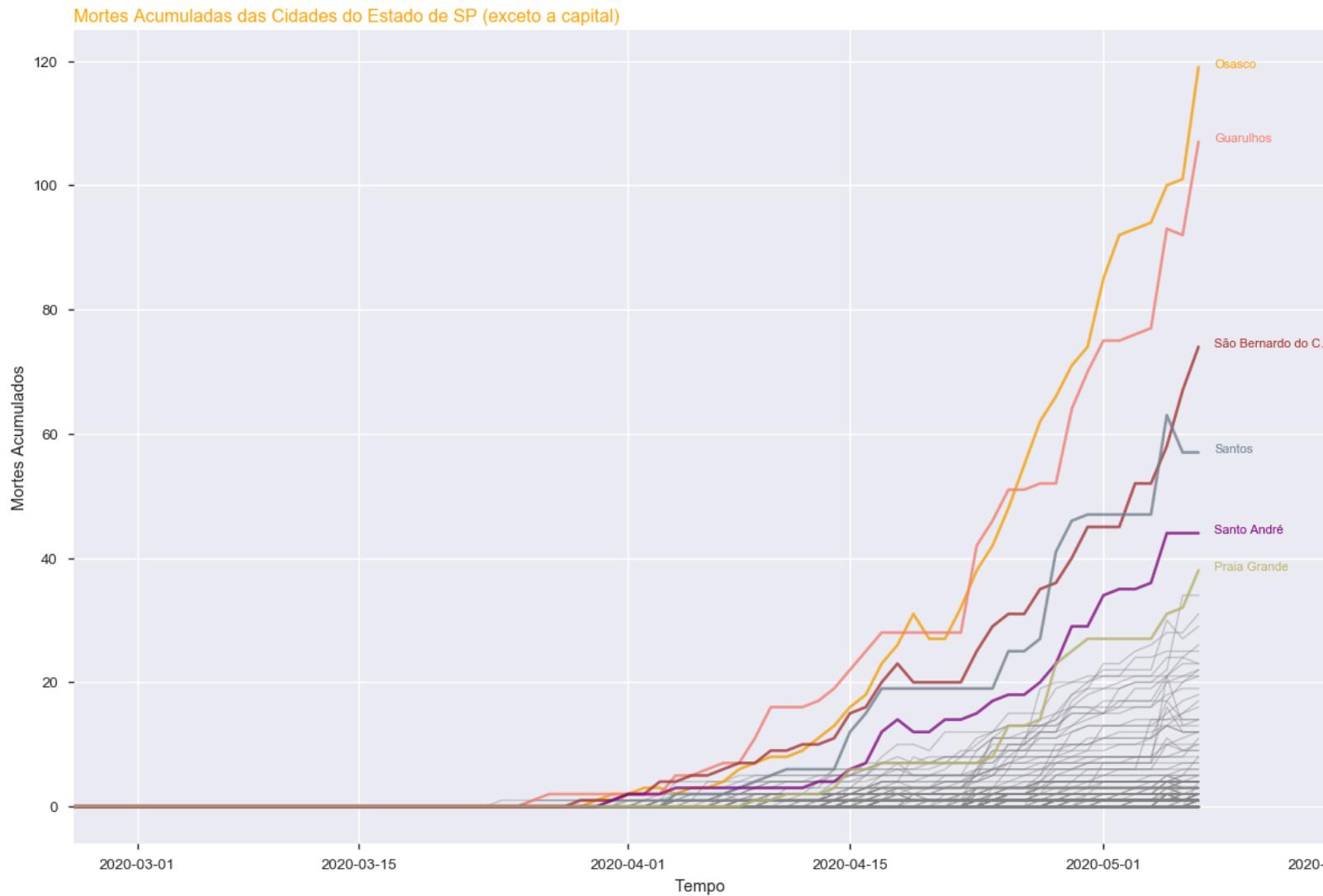
# 4. Análise Exploratória de Dados

RAIO-X DA BASE | ANÁLISE MULTIVARIADA

47

## Mortes Acumuladas por Cidade e por Data

- Praia Grande toma o lugar de Campinas, no Top 6 mortes acumuladas do Estado menos a capital.



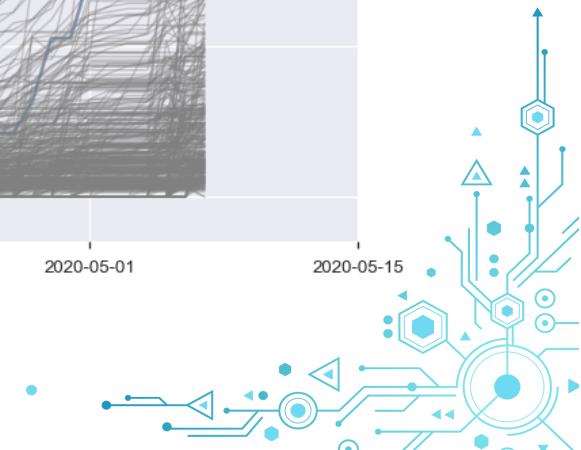
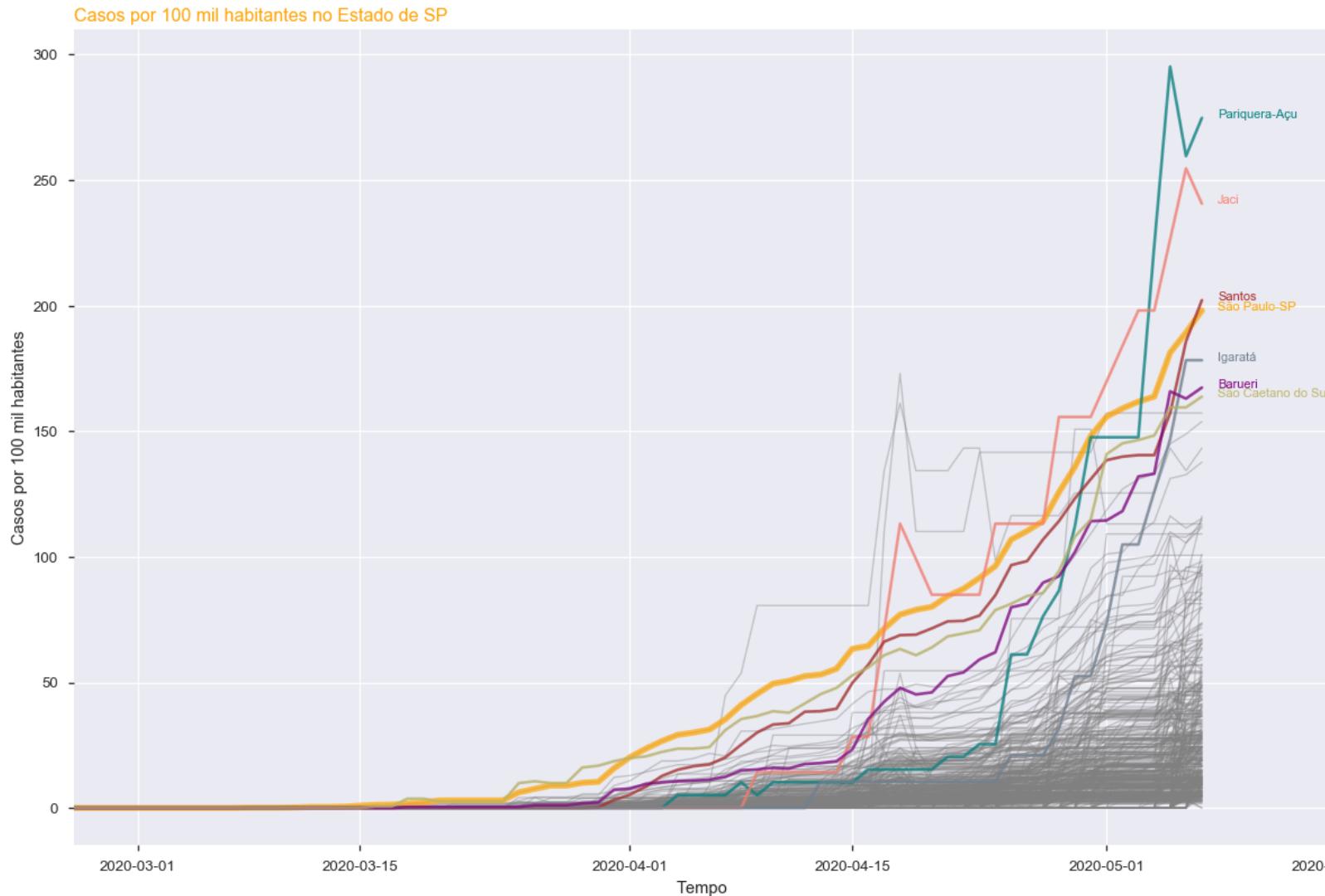
## 4. Análise Exploratória de Dados

RAIO-X DA BASE | ANÁLISE MULTIVARIADA

48

### Casos por 100 mil habitantes por Cidade e por Data

- Aqui, vemos as curvas nas cidades menores tomarem a dianteira, já que cada caso tem maior peso proporcional à população. Ainda assim, é de espantar as curvas de Santos e São Paulo. Também vale destacar como Barueri e São Caetano do Sul, muito próximas à capital, possuem curvas muito parecidas.

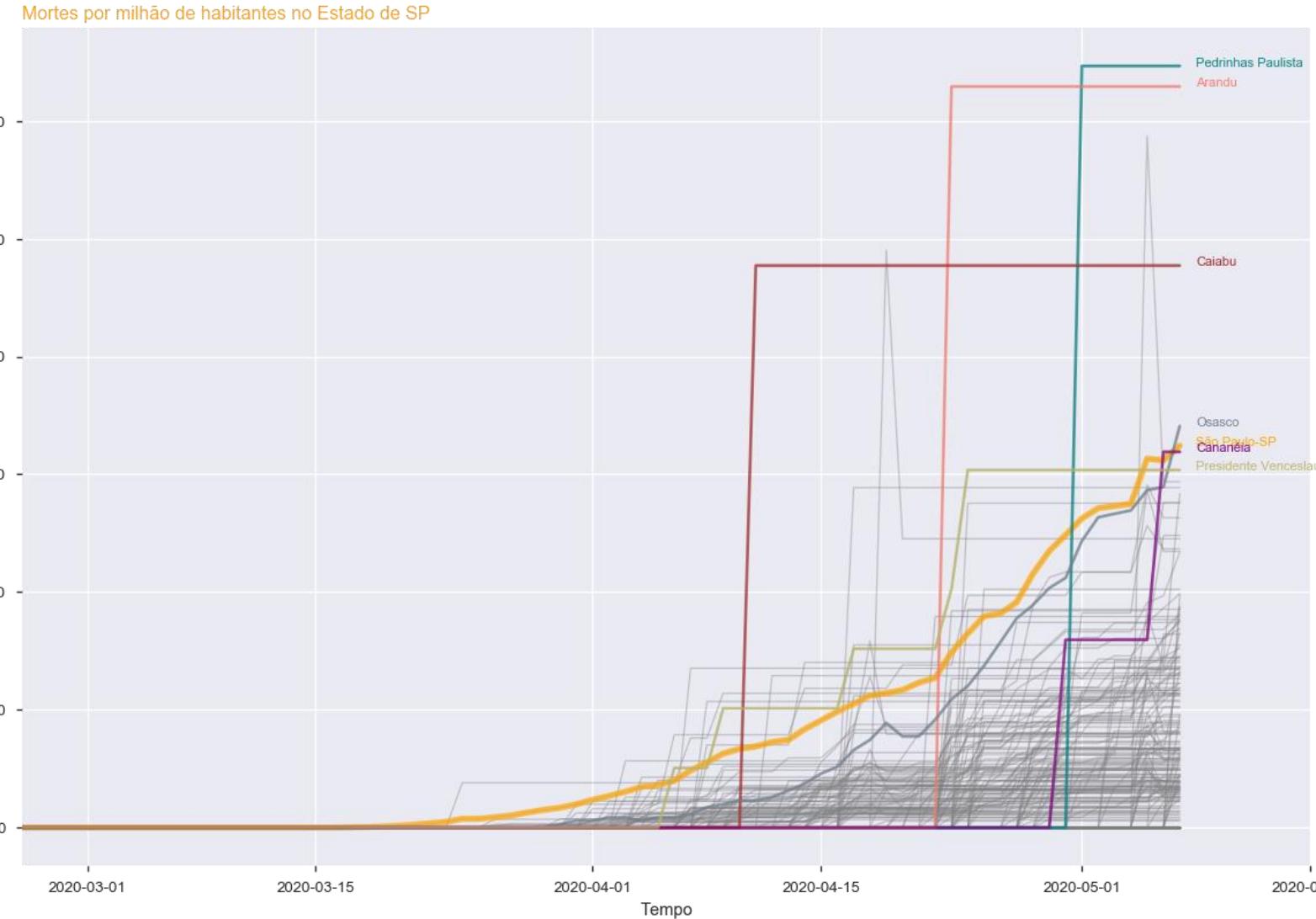


## 4. Análise Exploratória de Dados

RAIO-X DA BASE | ANÁLISE MULTIVARIADA

### Mortes por milhão de habitantes por Cidade e por Data

- Aqui, vemos as curvas nas cidades menores tomarem a dianteira, já que cada morte tem maior peso proporcional à população. Vale destacar as curvas de Osasco e São Paulo, cidades muito próximas, com comportamento muito parecido. O comportamento das curvas das cidades que possuem picos e estabilizam num patamar abaixo demonstram correção de registro de óbito.

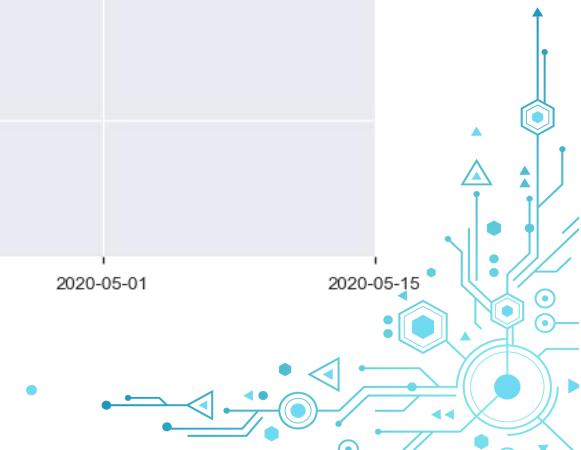
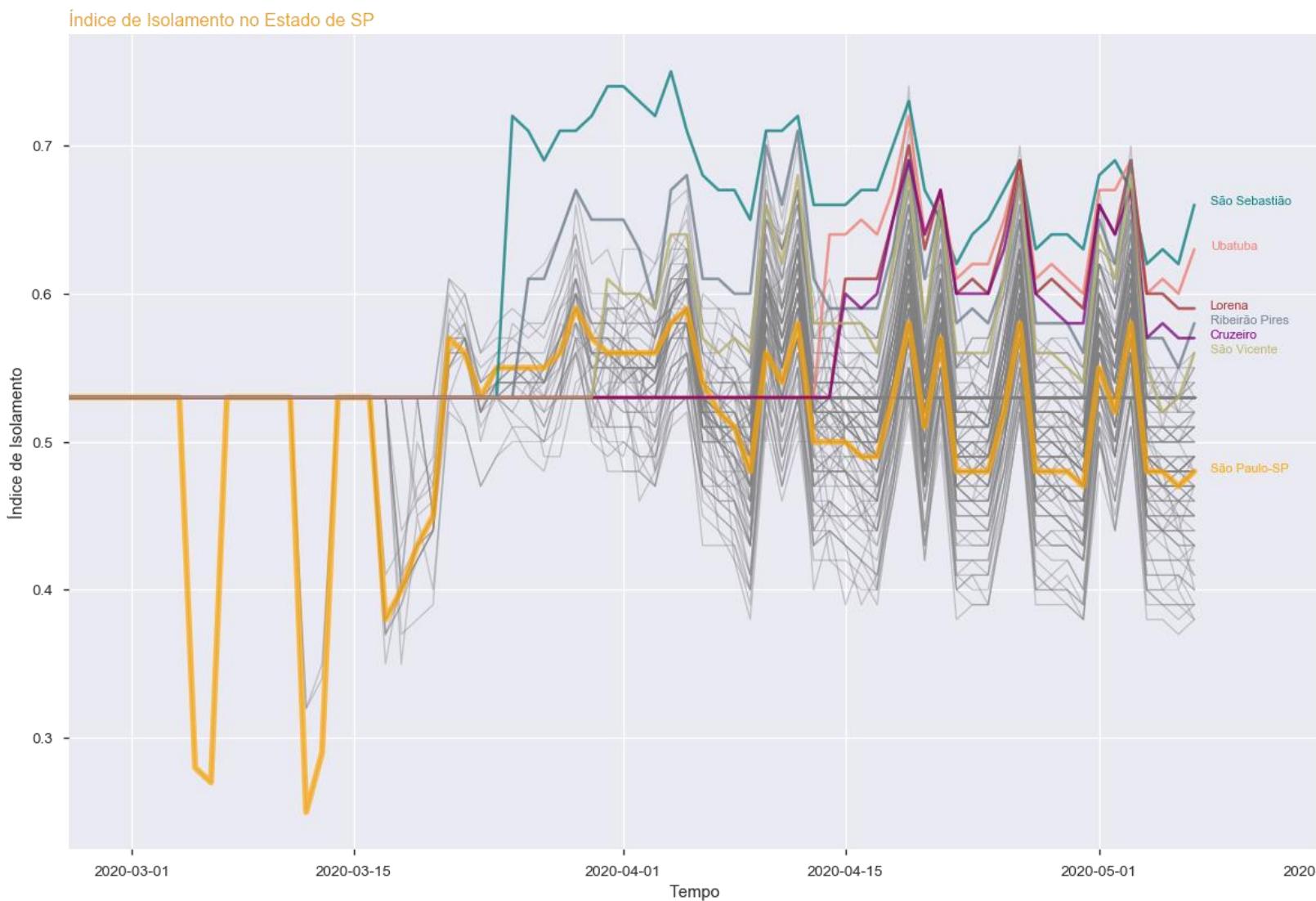


# 4. Análise Exploratória de Dados

RAIO-X DA BASE | ANÁLISE MULTIVARIADA

## Índice de Isolamento por Cidade e por Data

- De acordo com especialistas, a marca ideal do índice de isolamento para impedir a circulação do vírus é 70%. Vemos que pouquíssimas cidades do estado conseguiram superar a marca (São Sebastião, Ubatuba, Ribeirão Pires e Lorena), e por poucos dias. São Sebastião esteve por bastante tempo acima dos 70% devido ao bloqueio a turistas nas rodovias.



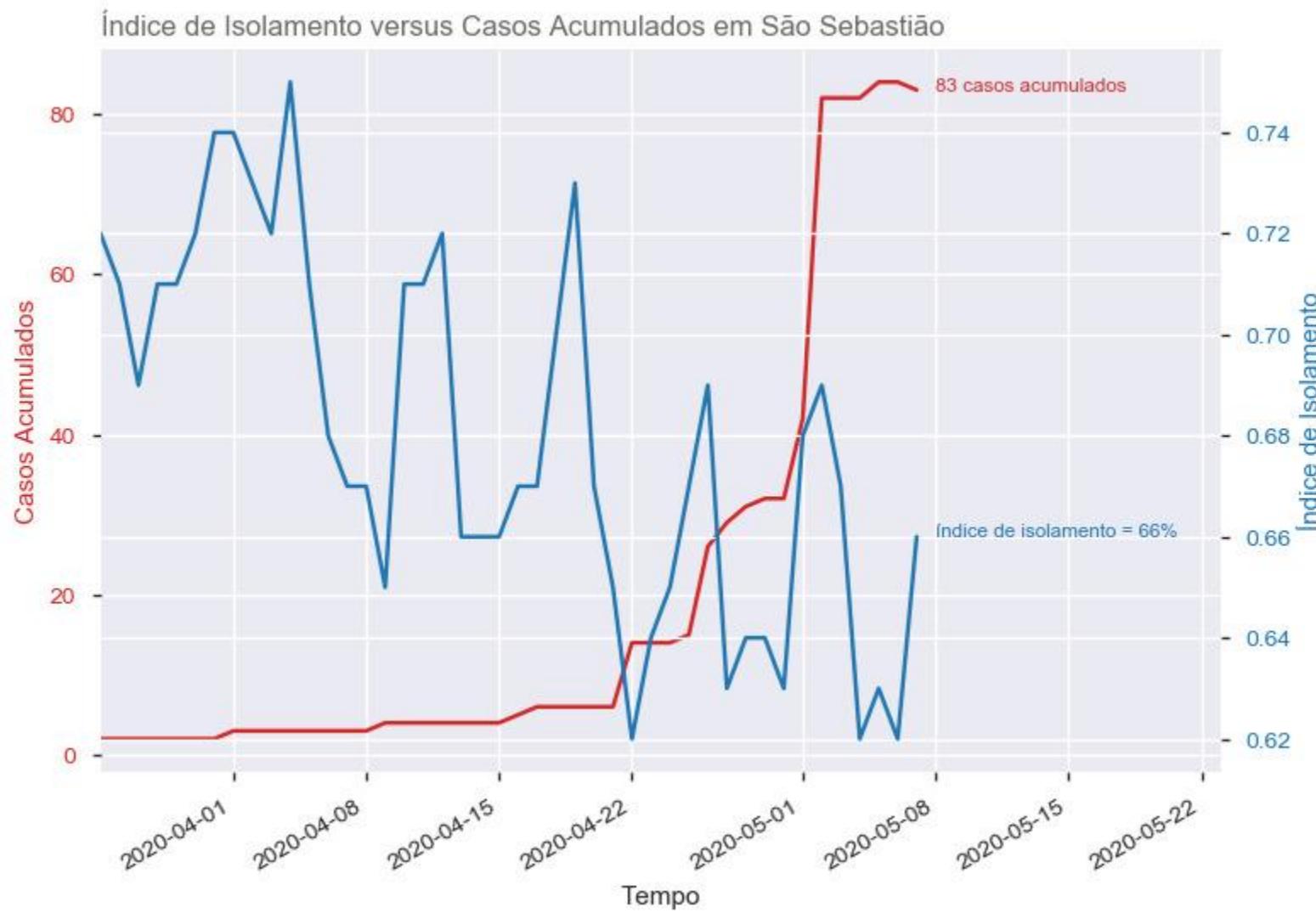
## 4. Análise Exploratória de Dados

RAIO-X DA BASE | ANÁLISE MULTIVARIADA

51

### Índice de Isolamento por Cidade e por Data

- Pode ser coincidência, mas aqui vemos São Sebastião, das raras cidades de SP que ficou algum tempo com índice de isolamento acima de 70% e baixo nº de casos de COVID-19. Quando o índice baixa desse patamar, a curva de casos sobe.



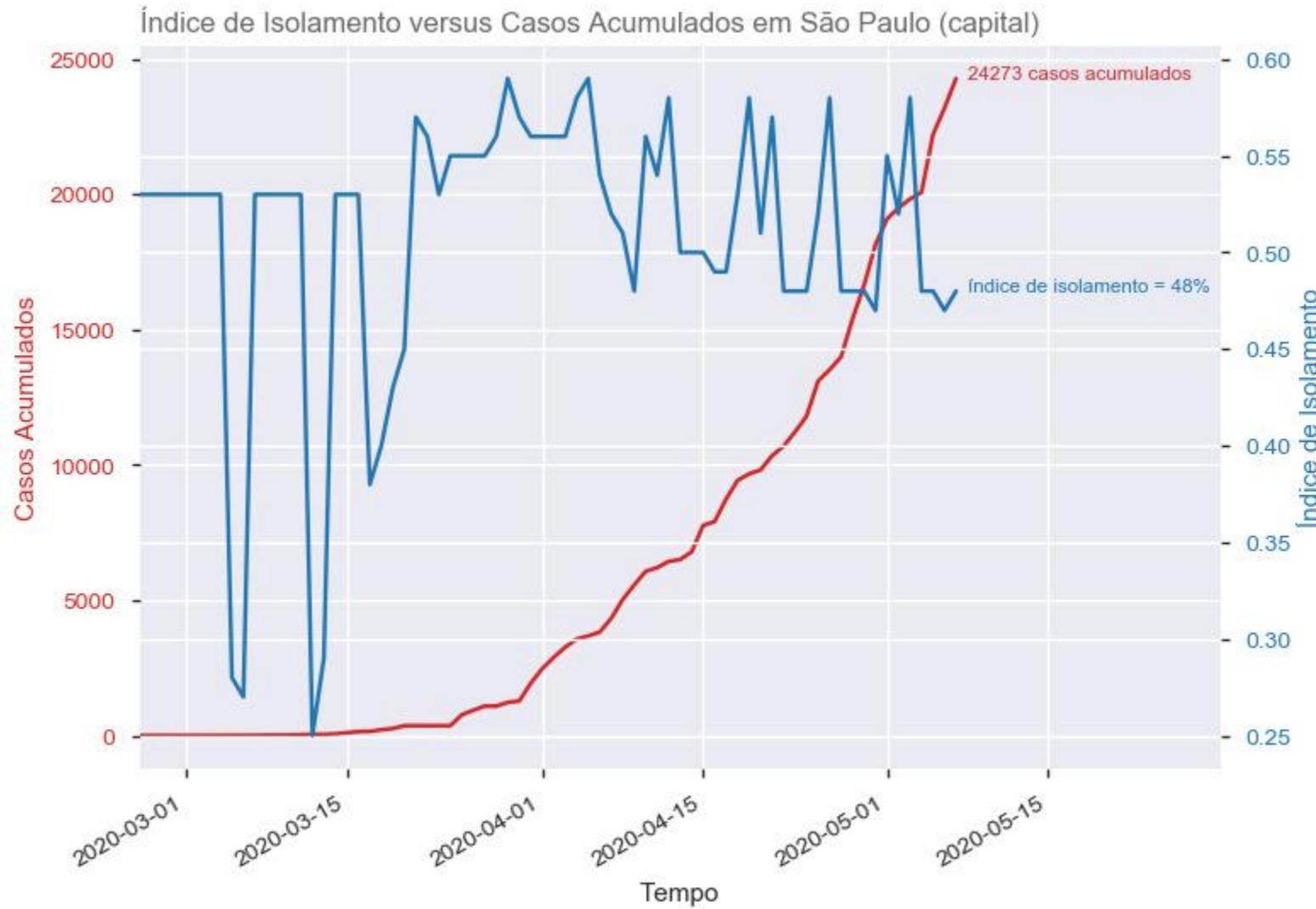
## 4. Análise Exploratória de Dados

RAIO-X DA BASE | ANÁLISE MULTIVARIADA

52

### Índice de Isolamento por Cidade e por Data

- E aqui, temo a capital São Paulo, que nunca atingiu 70% de índice de isolamento.



## 4. Análise Exploratória de Dados

RAIO-X DA BASE | ANÁLISE MULTIVARIADA

53

### Mapa de Casos Acumulados do Estado de São Paulo

- Situação em 01/03/2020
- Casos = pontos vermelhos



## 4. Análise Exploratória de Dados

RAIO-X DA BASE | ANÁLISE MULTIVARIADA

54

### Mapa de Casos Acumulados do Estado de São Paulo

- Situação em 15/03/2020
- Casos = pontos vermelhos



## 4. Análise Exploratória de Dados

RAIO-X DA BASE | ANÁLISE MULTIVARIADA

55

### Mapa de Casos Acumulados do Estado de São Paulo

- Situação em 01/04/2020
- Casos = pontos vermelhos



## 4. Análise Exploratória de Dados

RAIO-X DA BASE | ANÁLISE MULTIVARIADA

56

### Mapa de Casos Acumulados do Estado de São Paulo

- Situação em 15/04/2020
- Casos = pontos vermelhos



# 4. Análise Exploratória de Dados

57

RAIO-X DA BASE | ANÁLISE MULTIVARIADA

## Mapa de Casos Acumulados do Estado de São Paulo

- Situação em 01/05/2020
- Casos = pontos vermelhos



## 4. Análise Exploratória de Dados

RAIO-X DA BASE | ANÁLISE MULTIVARIADA

58

### Mapa de Casos Acumulados do Estado de São Paulo

- Situação em 07/05/2020
- Casos = pontos vermelhos



## 4. Análise Exploratória de Dados

RAIO-X DA BASE | ANÁLISE MULTIVARIADA

59

### Mapa de Mortes Acumuladas do Estado de São Paulo

- Situação em 01/03/2020
- Mortes = pontos verdes



# 4. Análise Exploratória de Dados

RAIO-X DA BASE | ANÁLISE MULTIVARIADA

60

## Mapa de Mortes Acumuladas do Estado de São Paulo

- Situação em 15/03/2020
- Mortes = pontos verdes



## 4. Análise Exploratória de Dados

61

RAIO-X DA BASE | ANÁLISE MULTIVARIADA

### Mapa de Mortes Acumuladas do Estado de São Paulo

- Situação em 01/04/2020
- Mortes = pontos verdes



## 4. Análise Exploratória de Dados

RAIO-X DA BASE | ANÁLISE MULTIVARIADA

62

### Mapa de Mortes Acumuladas do Estado de São Paulo

- Situação em 15/04/2020
- Mortes = pontos verdes



## 4. Análise Exploratória de Dados

RAIO-X DA BASE | ANÁLISE MULTIVARIADA

63

### Mapa de Mortes Acumuladas do Estado de São Paulo

- Situação em 01/05/2020
- Mortes = pontos verdes



## 4. Análise Exploratória de Dados

RAIO-X DA BASE | ANÁLISE MULTIVARIADA

64

### Mapa de Mortes Acumuladas do Estado de São Paulo

- Situação em 07/05/2020
- Mortes = pontos verdes



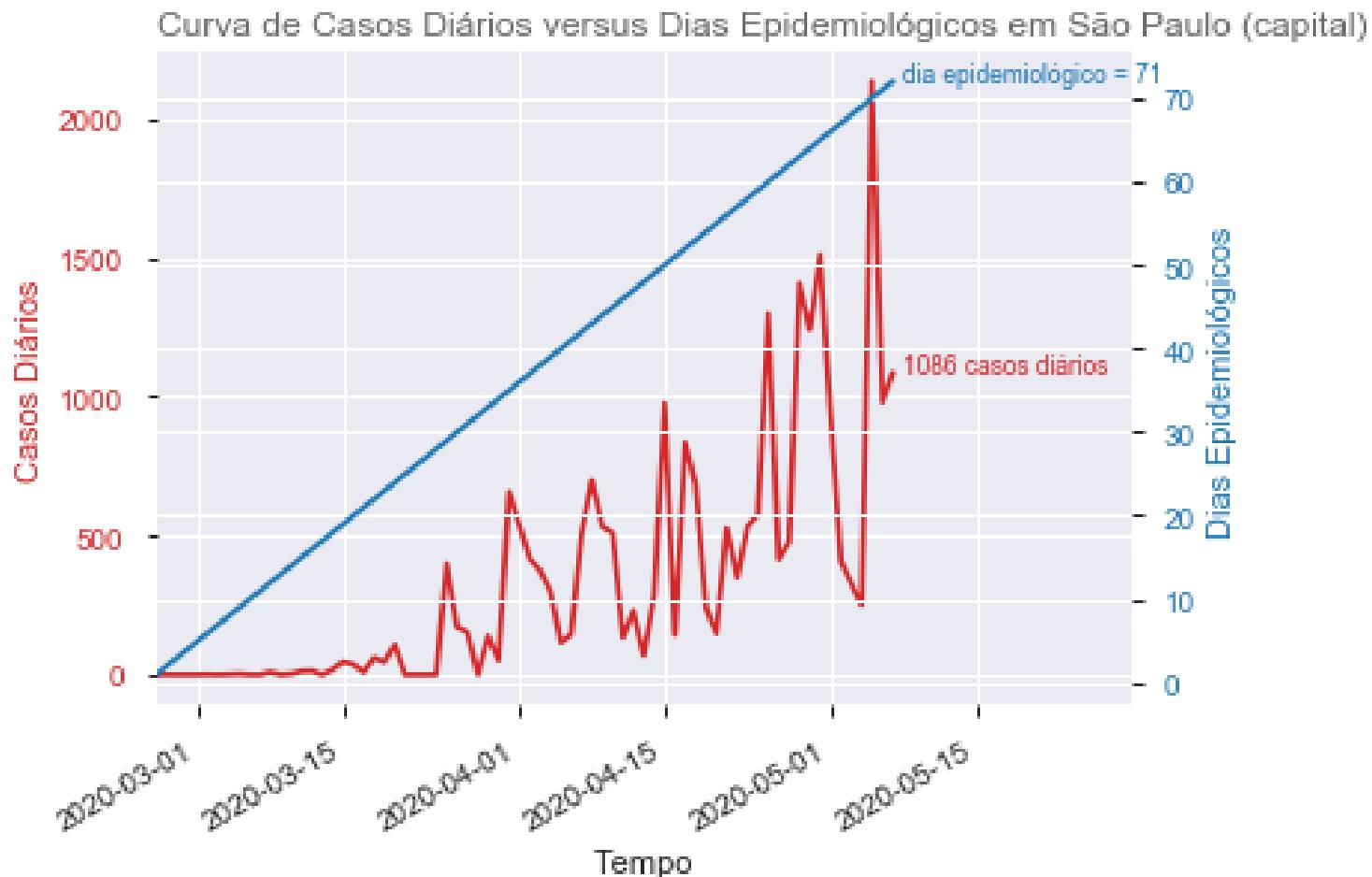
## 4. Análise Exploratória de Dados

RAIO-X DA BASE | ANÁLISE MULTIVARIADA

65

### Casos Diários por Dia Epidemiológico na Capital

- Neste recorte da capital, vemos como a tendência é que os casos diários tenham uma curva crescente conforme os dias epidemiológicos se acumulam.



# 4. Análise Exploratória de Dados

RAIO-X DA BASE | ANÁLISE MULTIVARIADA

## Correlação de Variáveis



# 4. Análise Exploratória de Dados

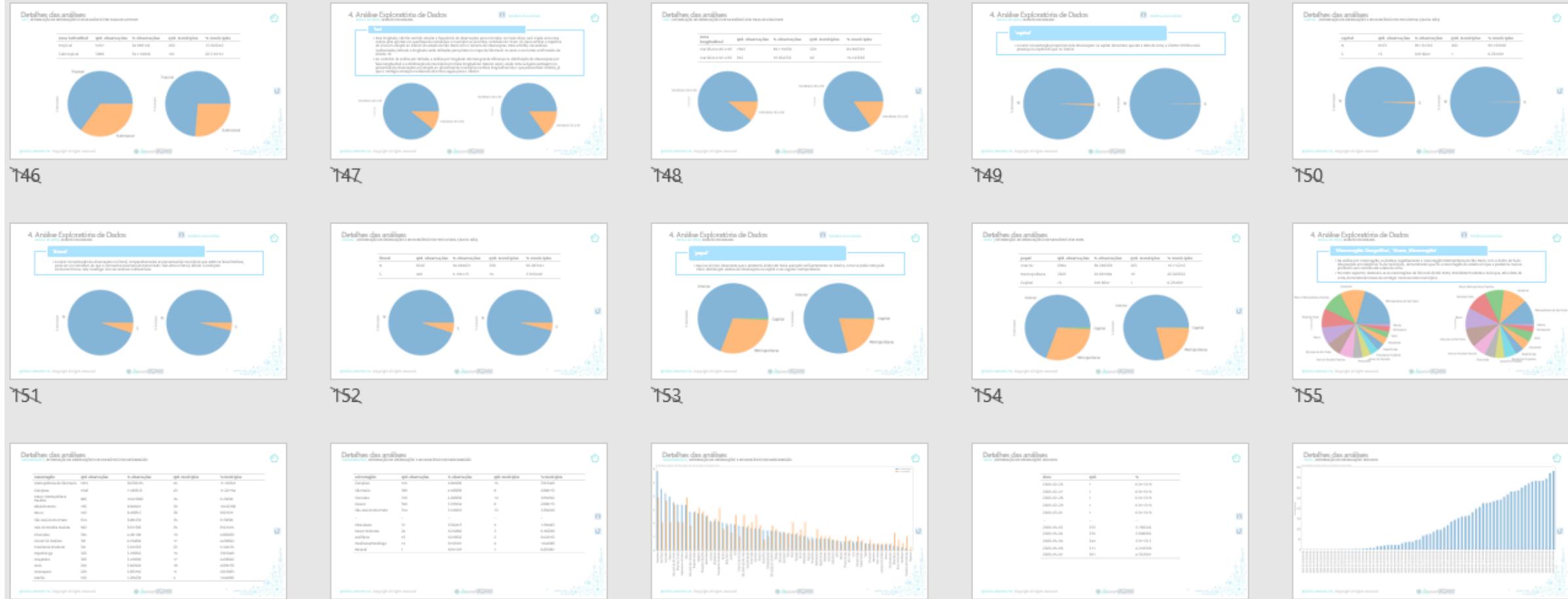
RAIO-X DA BASE | ANÁLISES COMPLEMENTARES

67

A análise das demais variáveis encontra-se como anexo:

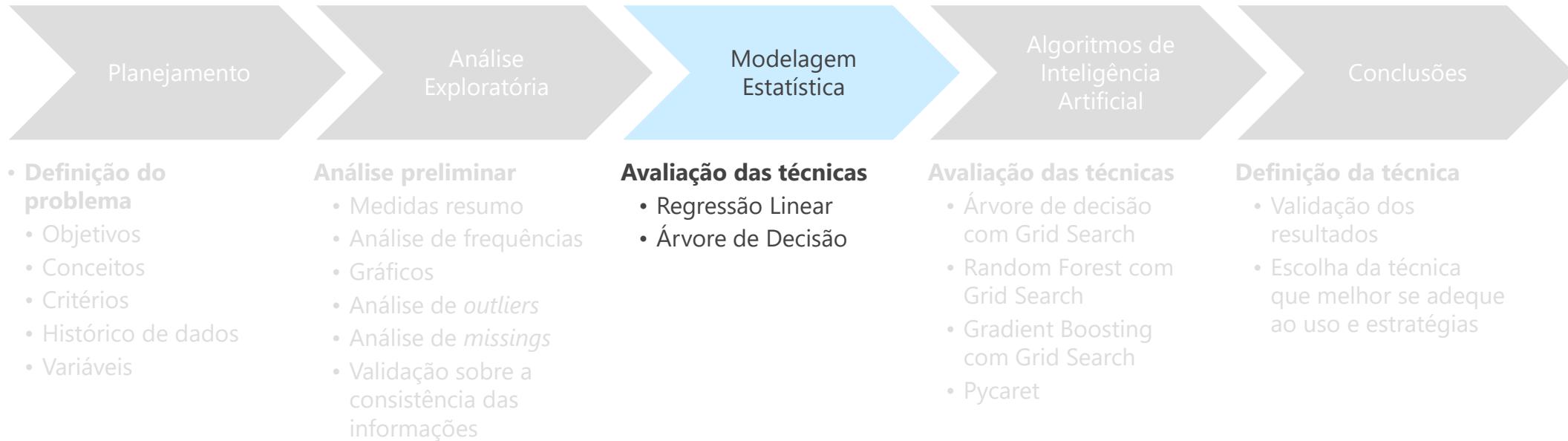


Detalhes das análises



# Metodologia de Análise de Dados

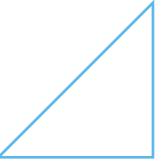
68



# 5. Modelagem com Estatística Tradicional

FEATURE ENGINEERING | PREPARAÇÃO DA ABT

69



## Exclusão de Variáveis:

- Foram excluídas as variáveis cod7d e cidade (ficando apenas com munuf para identificar o município), casos diários e casos por 100 mil habitantes (ficando apenas com casos acumulados), mortes diárias e mortes por milhão de habitantes (ficando apenas com casos acumulados), os totais de leitos SUS e Não SUS, o código da microrregião geográfica e seu nome (pois tem alta correlação com mesorregião geográfica e já temos o município como dado mais granular) e o código da mesorregião geográfica (ficando apenas com o nome da mesorregião).

## Criação de Novas Variáveis

- Dia do ano, em substituição à data.
- Lags para Casos e Mortes Acumuladas: uso de dados de 14 dias atrás para ajudar na previsão de hoje.
- Dummies para categóricas com baixa cardinalidade ('zona geográfica','faixa meridional','capital', 'litoral', 'papel', 'faixa\_pop').
- Label encoder para categóricas de alta cardinalidade ('munuf', Nome\_Mesorregião).



# 5. Modelagem com Estatística Tradicional

FEATURE ENGINEERING | BASELINE E VALIDAÇÃO

70

Vamos estabelecer que a base para comparação da previsão serão os casos e mortes do dia anterior. Ou seja, se prevíssemos que os casos e mortes de hoje serão iguais aos casos e mortes do dia anterior, o quanto estaríamos errando? O objetivo do modelo será superar esse erro médio.

## Métricas para avaliar o modelo

Para analisar se as previsões estão com os valores próximos dos dados reais deve-se fazer a medição do erro, o erro (ou resíduo) neste caso é basicamente  $Y_{real} - Y_{prev}$ .

Avalia-se o erro nos dados de treino para verificar se o modelo tem boa assertividade, e valida-se o modelo verificando o erro nos dados de teste (dados que não foram "vistos" pelo modelo).



# 5. Modelagem com Estatística Tradicional

BASELINE E VALIDAÇÃO | MÉTRICAS DE AVALIAÇÃO

71

## Principais métricas para avaliar modelos de séries temporais:

### Mean Forecast Error — (Erro Médio da Previsão ou Viés)

- Média dos erros da série avaliada.
- Essa métrica sugere que o modelo tende a fazer previsões acima do real (erros negativos) ou abaixo do real (erros positivos).

### MAE — Mean Absolute Error — (Erro Médio Absoluto)

- Muito semelhante ao viés. A única diferença é que o erro com valor negativo é transformado em positivo antes do cálculo da média.
- Uso em séries temporais: evita que o erro negativo cancele o positivo. Mostra o quanto a previsão está longe dos valores reais, independente se acima ou abaixo.

### MSE — Mean Squared Error — (Erro Quadrático Médio)

- Coloca mais peso nos erros maiores: cada valor individual do erro é elevado ao quadrado antes do cálculo da média. Muito sensível à outliers, coloca bastante peso nas previsões com erros mais significativos.
- Diferente do MAE e Viés, os valores do MSE estão em unidades quadráticas e não na unidade do modelo.

### RMSE — Root Mean Squared Error — (Erro Quadrático Médio da Raiz)

- Raiz quadrada do MSE: o erro volta a ter a unidade de medida do modelo.
- É muito usada em séries temporais porque é mais sensível à erros maiores devido ao processo de elevação ao quadrado que a originou.

### MAPE — Mean Absolute Percentage Error — (Erro Percentual Médio Absoluto)

- Erro é medido em termos percentuais e pode-se comparar o erro percentual do modelo de um objeto X com o erro percentual de um objeto Y.
- Cálculo: valor absoluto do erro dividido pelo resultado real, seguido do cálculo da média.



# 5. Modelagem com Estatística Tradicional

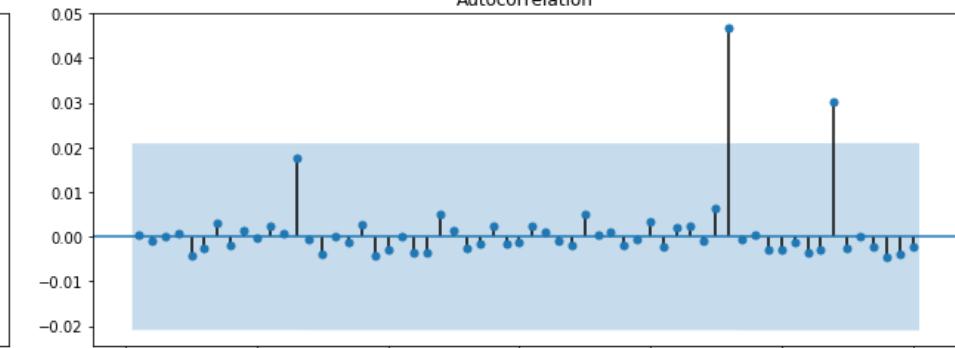
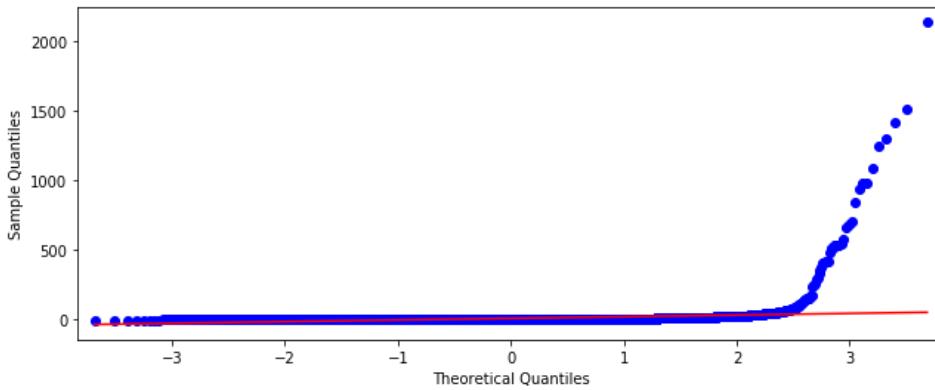
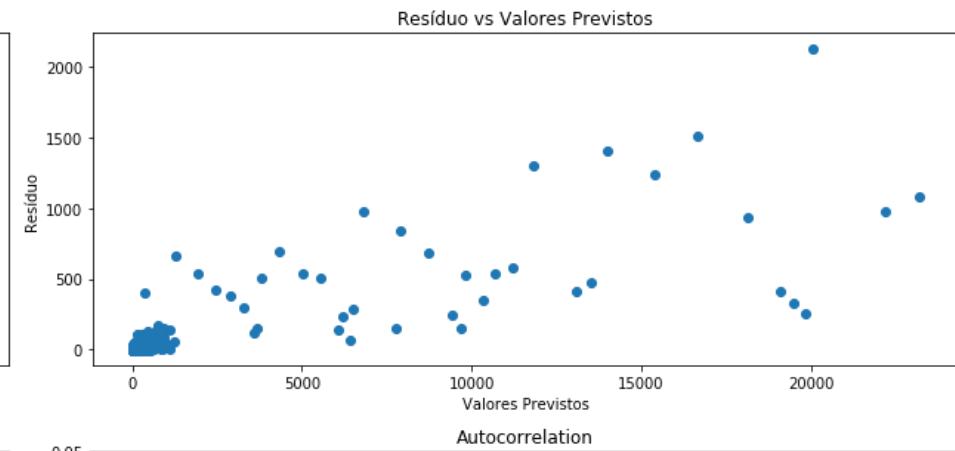
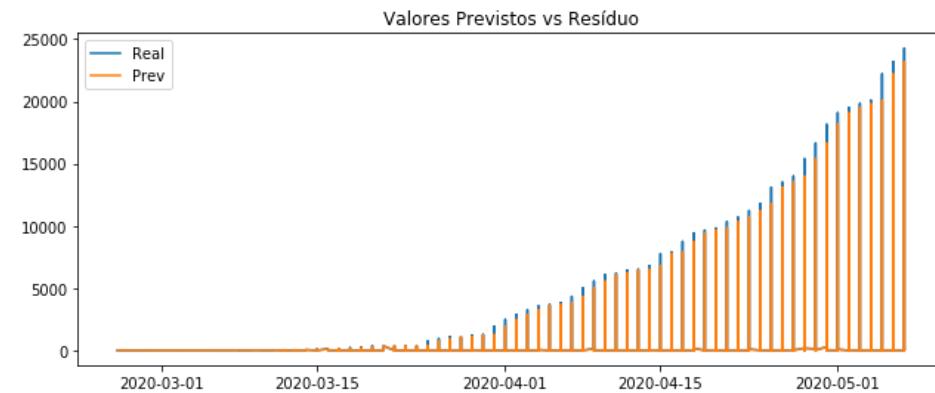
BASELINE E VALIDAÇÃO | CASOS ACUMULADOS

72

## Base Completa Baseline

VIÉS	4.54008
MSE	2534.639
RMSE	50.34520
MAE	4.65810
MAPE	11.77102

Os modelos terão a tarefa de superar a previsão com erro absoluto médio de 4.6 casos



# 5. Modelagem com Estatística Tradicional

BASELINE E VALIDAÇÃO | MORTES ACUMULADAS

73

## Base Completa

VIÉS 0.36453

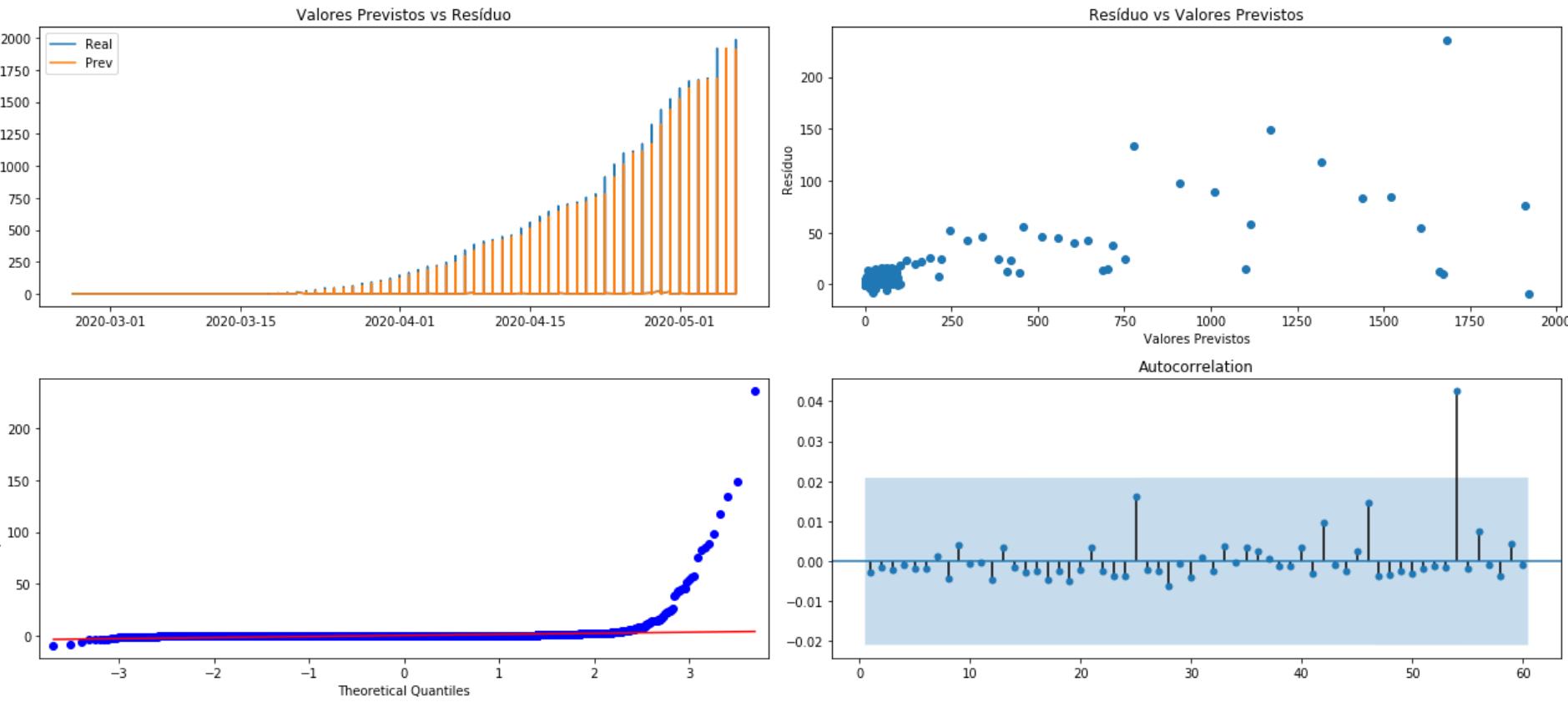
MSE 21.03945

RMSE 4.58688

MAE 0.38204

MAPE 9.12069

Os modelos terão a tarefa de superar a previsão com erro absoluto médio de 0.38 mortes

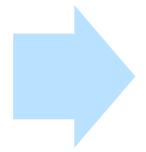


# 5. Modelagem com Estatística Tradicional

FEATURE ENGINEERING | PREPARAÇÃO DA ABT

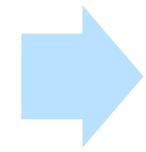
74

Separação da base em variáveis explicativas, target casos acumulados e target mortes acumuladas



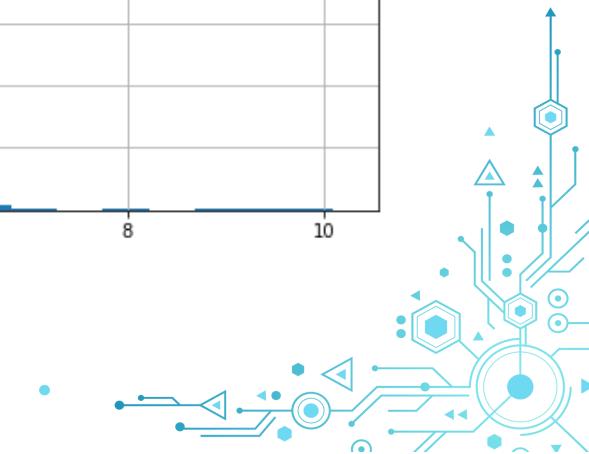
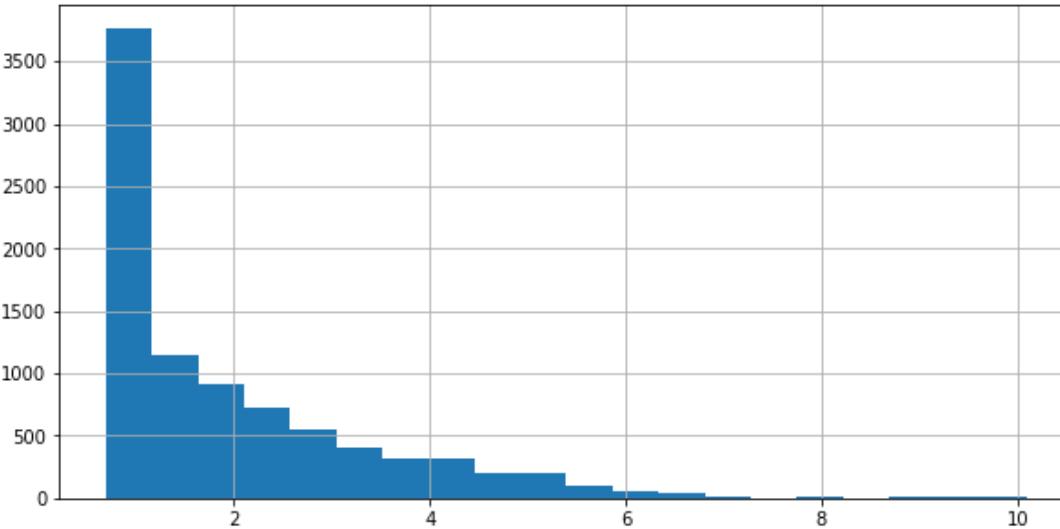
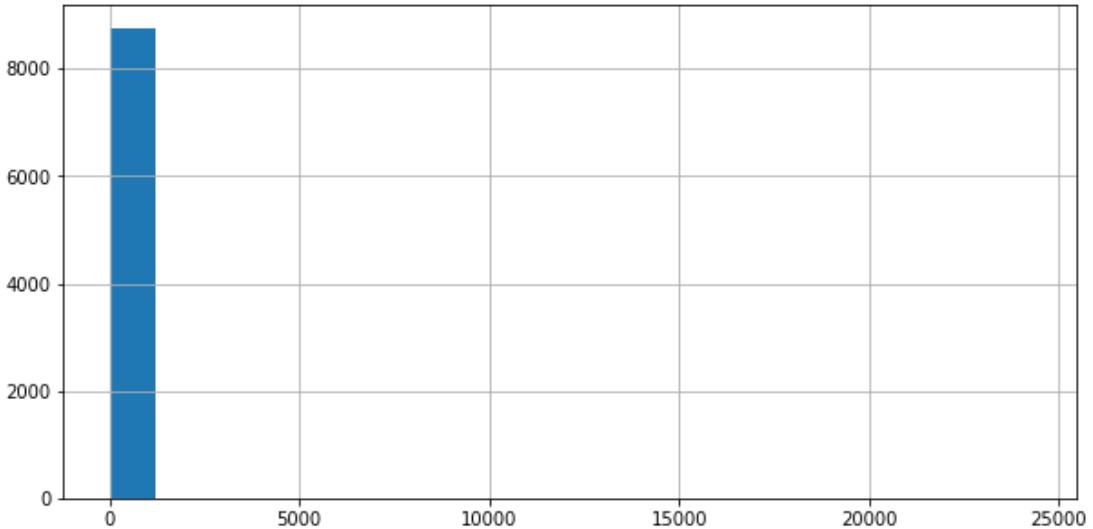
Normalização das variáveis explicativas

- Optamos por normalizar utilizando MinMax, para não haver interferência das ordens de grandeza das variáveis no modelo de regressão linear



Modificação da distribuição das targets

- Como a distribuição das targets está muito concentrada no zero, aplicamos uma função logarítmica para melhorar a performance do modelo.



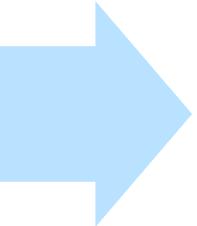
# 5. Modelagem com Estatística Tradicional

FEATURE ENGINEERING | SELEÇÃO DE VARIÁVEIS

75

Aplicamos 3 métodos de seleção de variáveis:

- Baseado em filtro ( $\chi^2$ )
- Baseado em wrapper (regressão linear)
- Baseado em método embarcado (random forest regressor)



Na comparação dos resultados de cada método, optamos por:

- **41** variáveis selecionadas em pelo menos 1 método para previsão de **casos acumulados** com regressão linear.
- **53** variáveis selecionadas em pelo menos 1 método para previsão de **casos acumulados** com árvore de decisão.
- **30** variáveis selecionadas em pelo menos 1 método para previsão de **mortes acumuladas** com regressão linear.
- **47** variáveis selecionadas em pelo menos 1 método para previsão de **mortes acumuladas** com árvore de decisão.



# 5. Modelagem com Estatística Tradicional

ELABORAÇÃO DE MODELOS | TÉCNICAS DE MODELAGEM DE SÉRIES TEMPORAIS

76

Estudando as técnicas de modelagem de séries temporais, selecionamos **2 métodos diferentes**:

No primeiro, fazemos uma **previsão em laço**, ou seja, analisamos os dados do 1º período para prever o 2º, aumentamos o 2º período ao 1º para prever o 3º e assim por diante.

No segundo, que chamaremos de **previsão tradicional**, dividiremos a base em teste e treino, sendo que a base teste será composta dos primeiros 80% da base e a base treino será composta dos últimos 20%, em ordem cronológica.

Como a curva de casos e mortes tende a crescer de forma exponencial, a regressão linear múltipla pode não dar conta sozinho de explicar as variáveis, pois seu resultado é uma reta. Por isso, realizamos testes com regressão polinomial, acrescentado termos elevados a potências, de forma a curvar a reta.



# 5.1. Modelagem com Estatística Tradicional

APLICAÇÃO DE MODELO | REGRESSÃO LINEAR

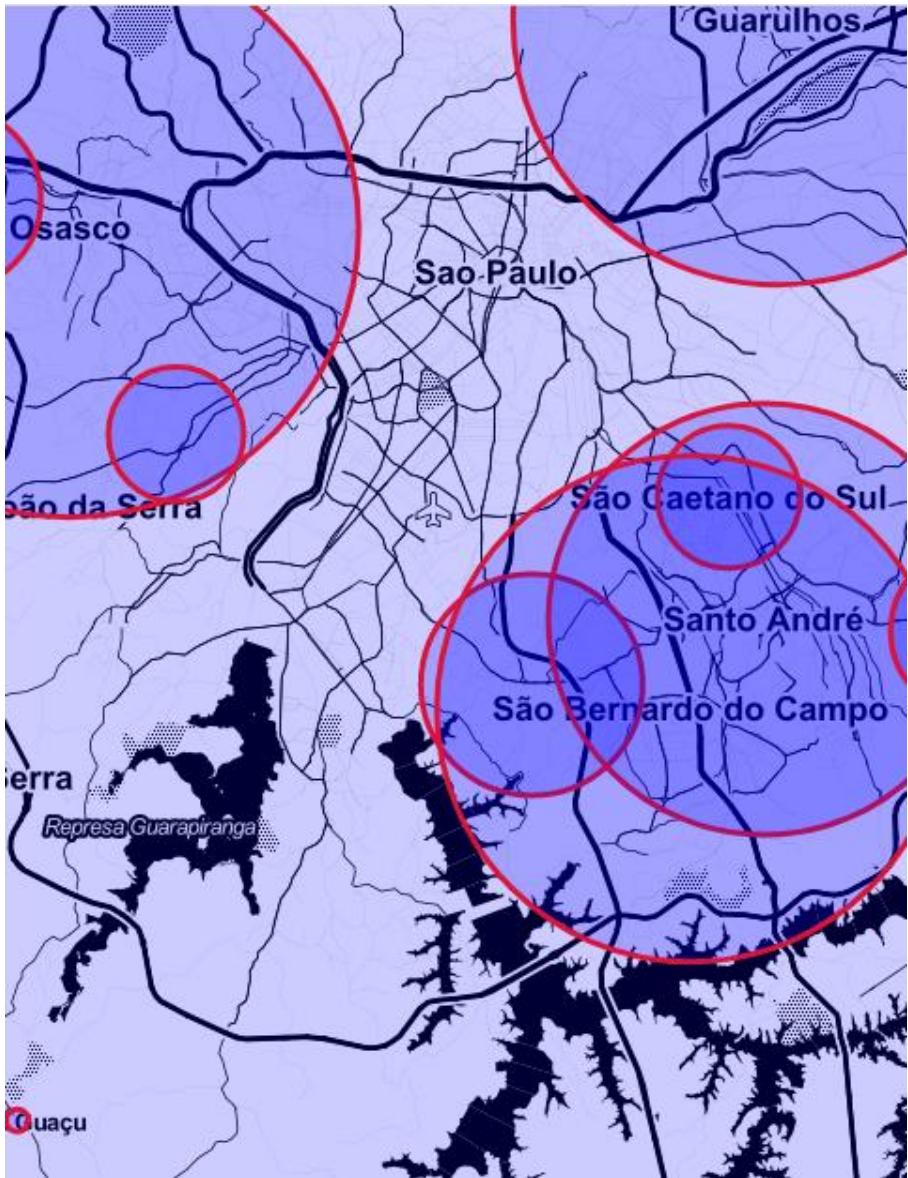
77

## REGRESSÃO LINEAR

## 5.1.1. Método de Previsão em Laço

REGRESSÃO LINEAR | CASOS ACUMULADOS

78



@2020 LABDATA FIA. Copyright all rights reserved.

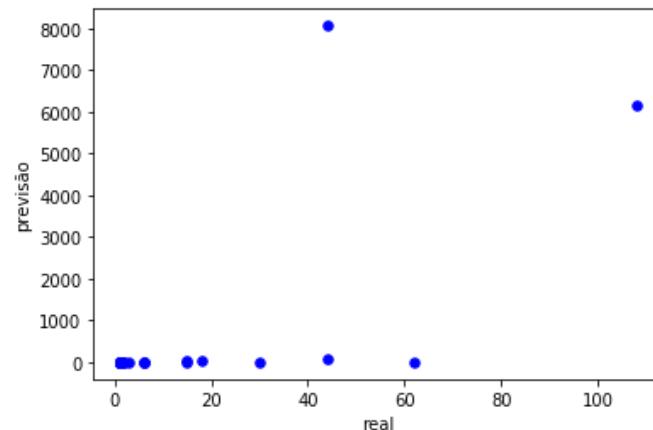
## 5.1.1. Método de Previsão em Laço

REGRESSÃO LINEAR | CASOS ACUMULADOS

79



Regressão linear sem polinômio:
• R <sup>2</sup> ajustado: 1.006 (BASE TREINO)
• Real   Previsto:
• Média: 11.51515   436.96970
• min: 1   1
• 25%: 1   1
• 50%: 1   2
• 75%: 6   6
• max: 108   8058
• sem previsões negativas

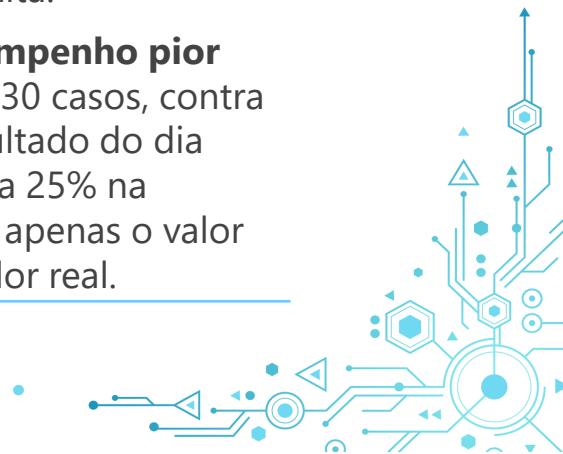


@2020 LABDATA FIA. Copyright all rights reserved.

BASE TESTE		
Indicadores sem polinômio para base de teste	BASELINE	Regressão Linear ScikitLearn (laço)
VIÉS	3.29412	-425.45455
MSE	84.88235	3046010.48485
RMSE	9.21316	1745.28235
MAE	3.29412	430.84848
MAPE	25.10127	761.49328

Com regressão linear múltipla **sem polinômio**, o laço funcionou até o dia 75, com indicadores de erros aumentando gradativamente. No dia 76, gerou previsão infinita.

Com a **base teste**, a regressão linear teve **desempenho pior que a baseline**. O erro médio absoluto foi de 430 casos, contra 3.2 casos na baseline (que apenas repete o resultado do dia anterior) e o MAPE aponta 761% de erro (contra 25% na baseline). Analisando os intervalos interquartis, apenas o valor máximo previsto ficou muito discrepante do valor real.

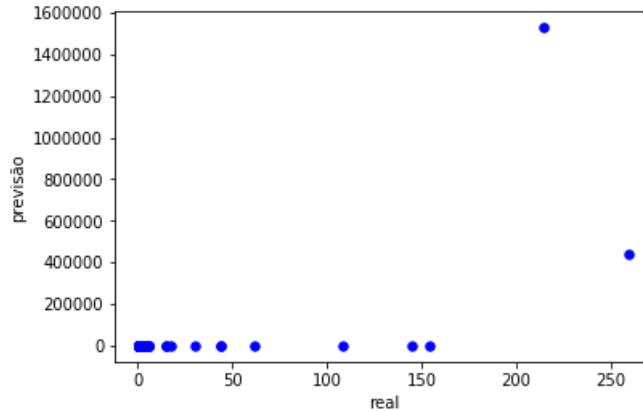


## 5.1.1. Método de Previsão em Laço

REGRESSÃO LINEAR | CASOS ACUMULADOS

80

Regressão linear c/ polinômio potência 2:
• R <sup>2</sup> ajustado: 1.0 (BASE TREINO)
• Real   Previsto:
• Média: 14.81707   24014.60976
• min: 1   -1
• 25%: 1   1
• 50%: 1   1
• 75%: 2   3.75
• max: 259   1527516
• 1 previsão negativa (-1 previsto contra 6 real)

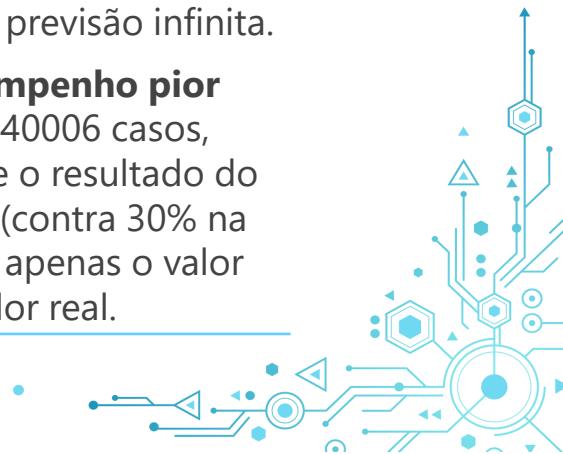


### BASE TESTE

Indicadores com polinômio com potência 2 para base de teste	BASELINE	Regressão Linear ScikitLearn (laço)
VIÉS	3.44578	-23999.79268
MSE	120.55422	30793810558.69512
RMSE	10.97972	175481.65305
MAE	3.44578	24006.20732
MAPE	30.48453	11066.67157

Com regressão linear múltipla **com polinômio de potência 2**, o laço funcionou até o **dia 79**, com indicadores de erros aumentando gradativamente. No dia 80, gerou previsão infinita.

Com a **base teste**, a regressão linear teve **desempenho pior que a baseline**. O erro médio absoluto foi de 240006 casos, contra 3.4 casos na baseline (que apenas repete o resultado do dia anterior) e o MAPE aponta 11066% de erro (contra 30% na baseline). Analisando os intervalos interquartis, apenas o valor máximo previsto ficou muito discrepante do valor real.



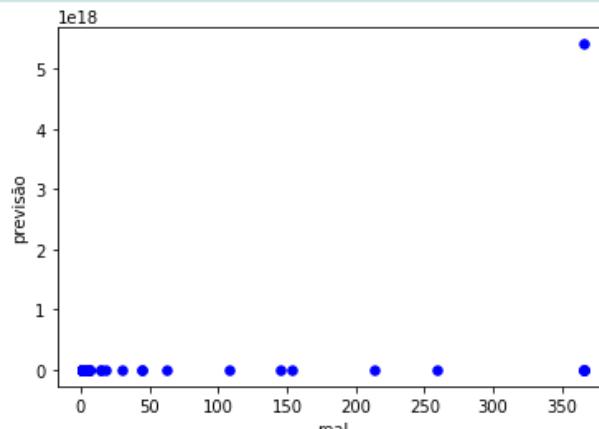
## 5.1.1. Método de Previsão em Laço

REGRESSÃO LINEAR | CASOS ACUMULADOS

81

### Regressão linear c/ polinômio potência 3:

- R<sup>2</sup> ajustado: 1.0 (BASE TREINO)
- Real | Previsto:
  - média: 16.54639 | 27909598283617284
  - min: 1 | -1
  - 25%: 1 | 1
  - 50%: 1 | 1
  - 75%: 2 | 2
  - max: 366 | 5414462067015961600
- 17 previsões negativas (-1 previsto contra 1, 2 e 366 reais)

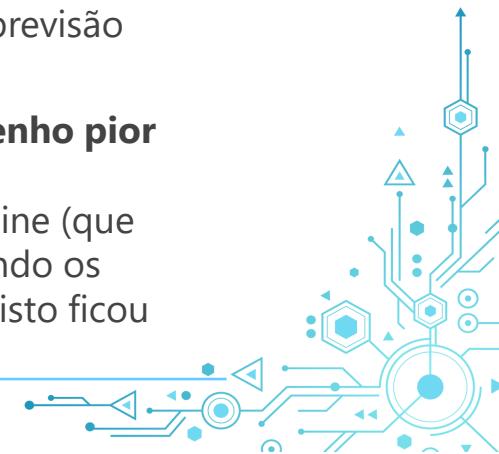


### BASE TESTE

Indicadores com polinômio com potência 3 para base de teste	BASELINE	Regressão Linear ScikitLearn (laço)
VIÉS	2.05641	-27909598283617280
MSE	110.09744	14395125276719640
RMSE	10.49273	119979686.93375
MAE	2.07692	27909598283617292
MAPE	16.92881	7625573302110736

Com regressão linear múltipla **com polinômio de potência 3**, o laço funcionou até o **dia 84**, com indicadores de erros aumentando exponencialmente. No dia 85, gerou previsão infinita.

Com a **base teste**, a regressão linear teve **desempenho pior que a baseline**. O erro médio absoluto foi de 27909598283617292 casos, contra 2 casos na baseline (que apenas repete o resultado do dia anterior). Analisando os intervalos interquartis, apenas o valor máximo previsto ficou muito discrepante do valor real.



## 5.1.1. Método de Previsão em Laço

REGRESSÃO LINEAR | CASOS ACUMULADOS

82

### Conclusões:

- o método de previsão em laço com regressão linear não funcionou para prever casos acumulados.
- há uma relação diretamente proporcional entre a potência usada nos polinômios e a quantidade de dias que o modelo consegue prever.
- há uma relação inversamente proporcional entre a quantidade de dias que o modelo consegue prever e o erro.



## 5.1.1. Método de Previsão em Laço

REGRESSÃO LINEAR | MORTES ACUMULADAS



83

PREVISÃO  
EM LAÇO

REGRESSÃO  
LINEAR

MORTES  
ACUMULADAS



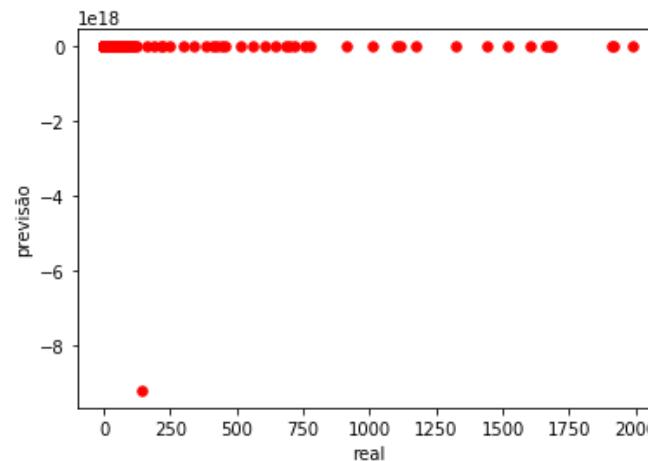
## 5.1.1. Método de Previsão em Laço

REGRESSÃO LINEAR | MORTES ACUMULADAS

84



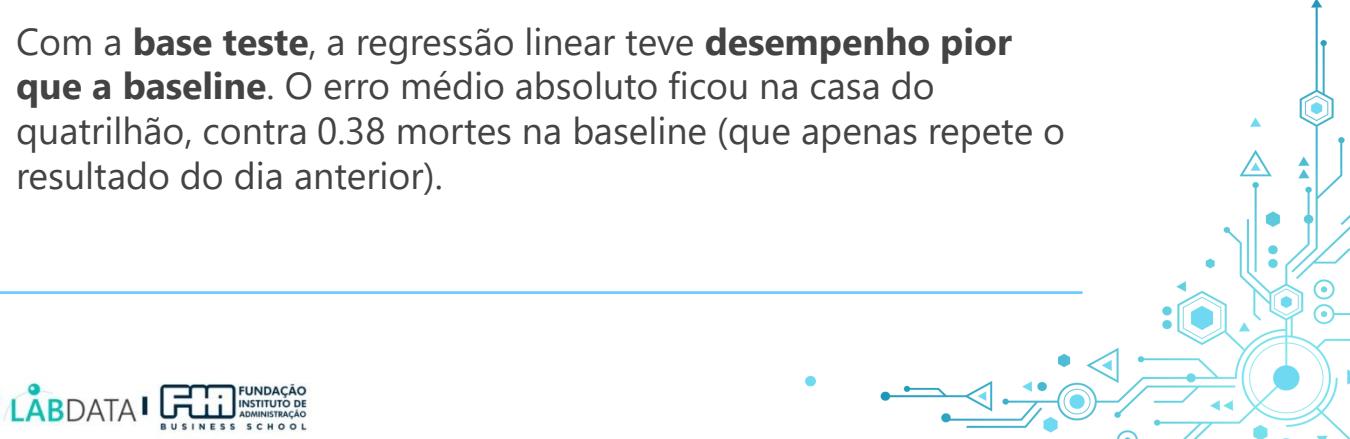
Regressão linear sem polinômio:
• R <sup>2</sup> ajustado: 0.3724 (BASE TREINO)
• Real   Previsto:
• Média: 5.53411   -1048821053610022.25
• min: 0   -9223372036854775808
• 25%: 0   0
• 50%: 0   0
• 75%: 1   0
• max: 1986   30144567005195
• 62 previsões negativas



BASE TESTE		
Indicadores sem polinômio para base de teste	BASELINE	Regressão Linear ScikitLearn (laço)
VIÉS	0.36453	-1048830080539344
MSE	21.03945	2590765068150100
RMSE	4.58688	50899558.62432
MAE	0.38204	1048830080539347
MAPE	nan	inf

Com regressão linear múltipla **sem polinômio**, o laço funcionou até o dia 128, com indicadores de erros aumentando absurdamente.

Com a **base teste**, a regressão linear teve **desempenho pior que a baseline**. O erro médio absoluto ficou na casa do quatrilhão, contra 0.38 mortes na baseline (que apenas repete o resultado do dia anterior).



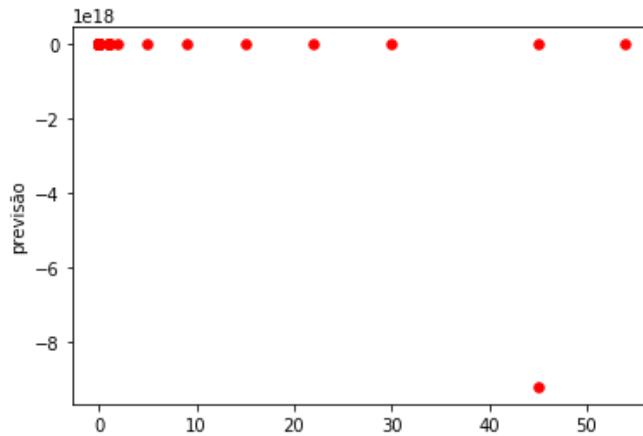
## 5.1.1. Método de Previsão em Laço

REGRESSÃO LINEAR | MORTES ACUMULADAS

85

### Regressão linear c/ polinômio potência 2:

- R<sup>2</sup> ajustado: 0.9923 (BASE TREINO)
- Real | Previsto:
  - Média: 0.89655 | -35338590174204476
  - min: 0 | -9223372036854775808
  - 25%: 0 | 0
  - 50%: 0 | 0
  - 75%: 0 | 0
  - max: 54 | 1387402183
- 7 previsões negativas

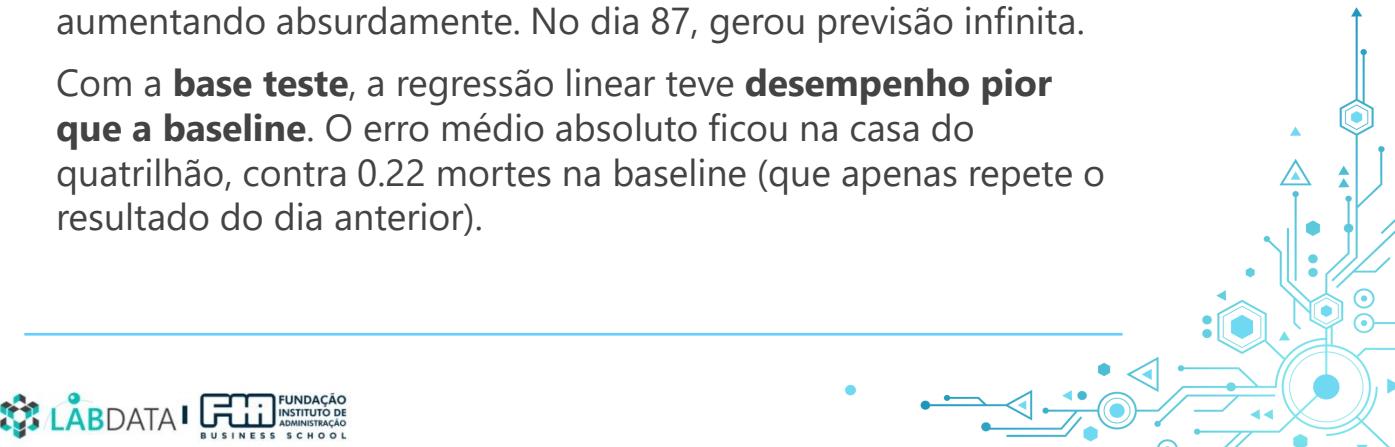


### BASE TESTE

Indicadores com polinômio com potência 2 para base de teste	BASELINE	Regressão Linear ScikitLearn (laço)
VIÉS	0.22137	-35338590184835952
MSE	1.85496	7375037136190711
RMSE	1.36197	85878036.40158
MAE	0.22137	35338590184835944
MAPE	nan	inf

Com regressão linear múltipla **com polinômio de potência 2**, o laço funcionou até o dia 86, com indicadores de erros aumentando absurdamente. No dia 87, gerou previsão infinita.

Com a **base teste**, a regressão linear teve **desempenho pior que a baseline**. O erro médio absoluto ficou na casa do quatrilhão, contra 0.22 mortes na baseline (que apenas repete o resultado do dia anterior).



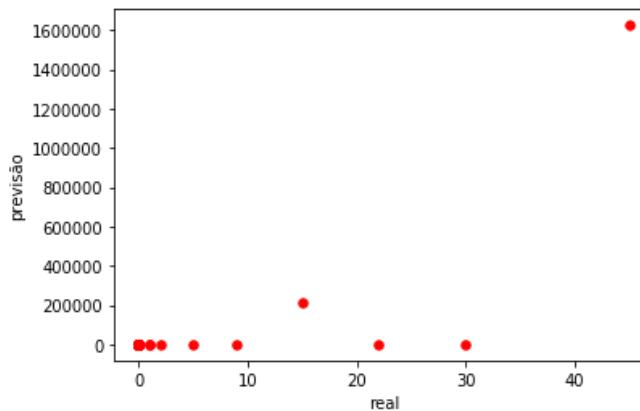
## 5.1.1. Método de Previsão em Laço

REGRESSÃO LINEAR | MORTES ACUMULADAS

86

### Regressão linear c/ polinômio potência 3:

- R<sup>2</sup> ajustado: 1.0 (BASE TREINO)
- Real | Previsto:
  - média: 0.67010 | 9475.02062
  - min: 0 | -1
  - 25%: 0 | 0
  - 50%: 0 | 0
  - 75%: 0 | 0
  - max: 45 | 1624446
- 2 previsões negativas (-1 previsto contra 22 e 30 reais)

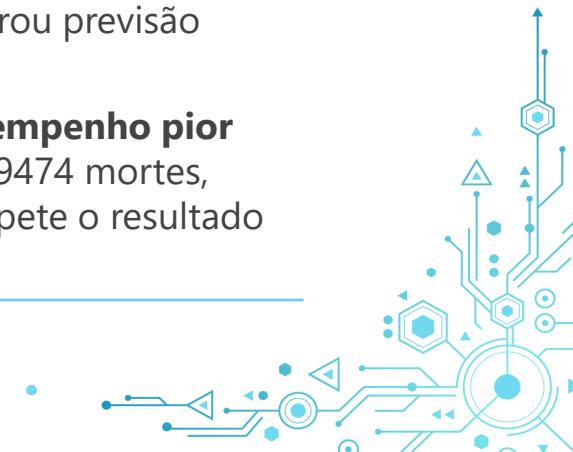


BASE TESTE

Indicadores com polinômio com potência 3 para base de teste	BASELINE	Regressão Linear ScikitLearn (laço)
VIÉS	0.23590	-9474.35052
MSE	2.06154	13836783728.71134
RMSE	1.43581	117629.85900
MAE	0.23590	9474.98969
MAPE	nan	559441.61616

Com regressão linear múltipla **com polinômio de potência 3**, o laço funcionou até o dia 84, com indicadores de erros aumentando exponencialmente. No dia 85, gerou previsão infinita.

Com a **base teste**, a regressão linear teve **desempenho pior que a baseline**. O erro médio absoluto foi de 9474 mortes, contra 0.23 mortes na baseline (que apenas repete o resultado do dia anterior).



## 5.1.1. Método de Previsão em Laço

REGRESSÃO LINEAR | MORTES ACUMULADAS

87

### Conclusões:

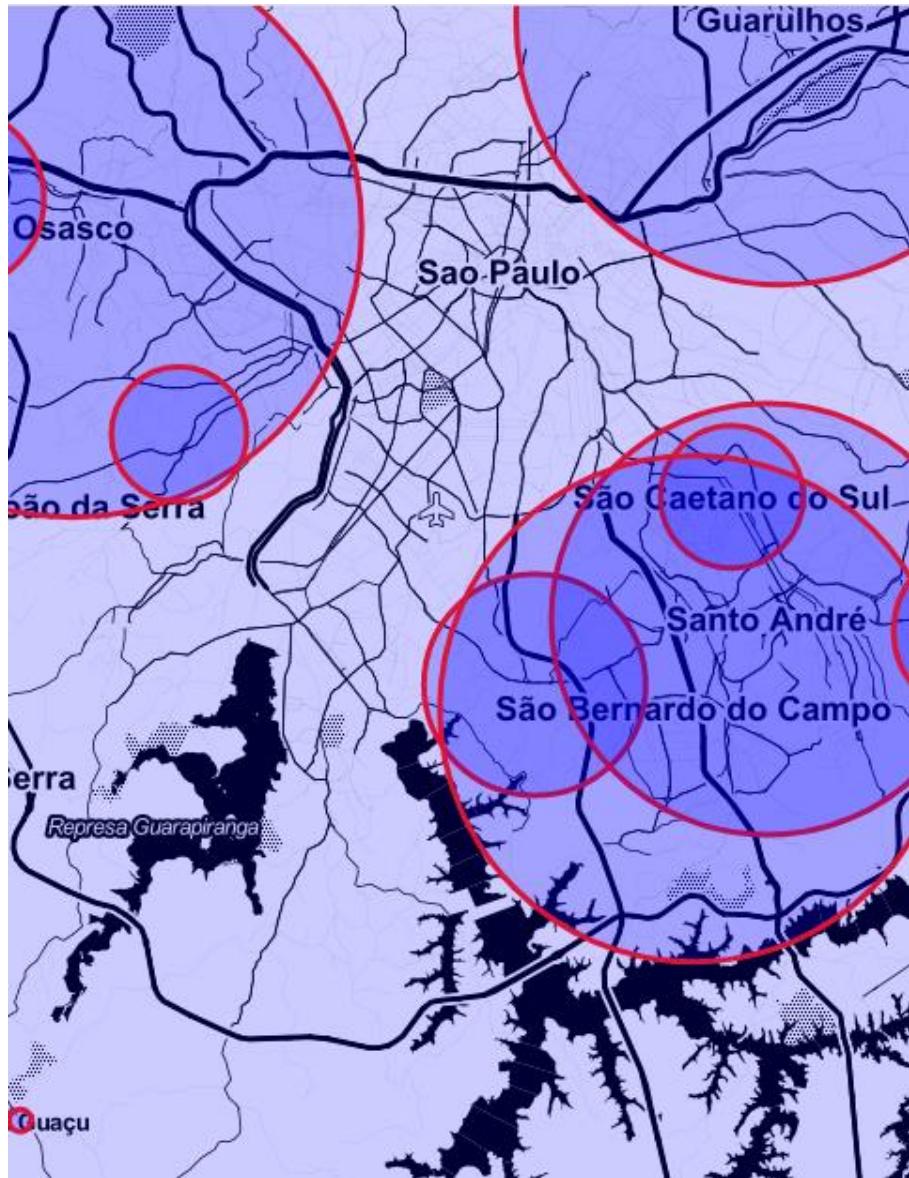
- o método de previsão em laço com regressão linear não funcionou para prever mortes acumuladas, tendo como resultado previsões absurdas.



## 5.1.2. Método de Previsão Tradicional

REGRESSÃO LINEAR | CASOS ACUMULADOS

88



@2020 LABDATA FIA. Copyright all rights reserved.



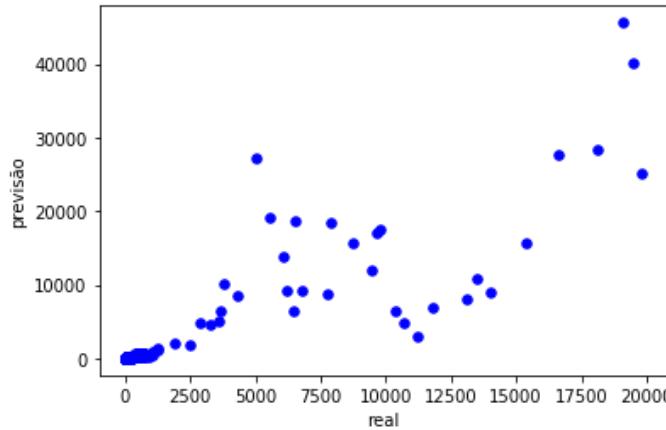
## 5.1.2. Método de Previsão Tradicional

REGRESSÃO LINEAR | CASOS ACUMULADOS

89

### BASE TREINO

Regressão linear sem polinômio:	
• R <sup>2</sup> ajustado:	0.7740060132065394
• Real   Previsto:	
• média:	65.09791   79.94527
• min:	1   -1
• 25%:	1   2
• 50%:	3   4
• 75%:	11   10
• max:	19822   45637
• 4 previsões negativas (-1 previsto contra 1 real)	



Indicadores sem polinômio para base de treino	BASELINE	Regressão Linear ScikitLearn
VIÉS	4.31545	-14.84737
MSE	2087.44337	386216.65318
RMSE	45.68855	621.46332
MAE	4.41214	40.32319
MAPE	11.90706	95.74353

Com a **base treino**, a regressão linear teve **desempenho pior que a baseline**. O erro médio absoluto foi de 40 casos, contra 4.4 casos na baseline (que apenas repete o resultado do dia anterior) e o MAPE aponta 95.7% de erro (contra 11.9% na baseline).

Analizando os intervalos interquartis, apenas o valor máximo previsto ficou muito discrepante do valor real.



## 5.1.2. Método de Previsão Tradicional

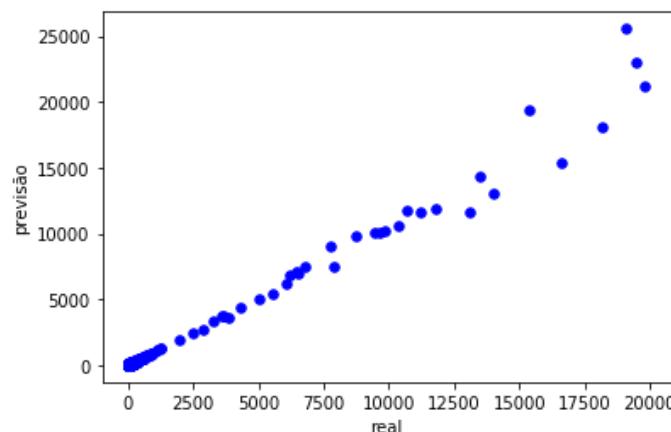
REGRESSÃO LINEAR | CASOS ACUMULADOS

90

BASE TREINO

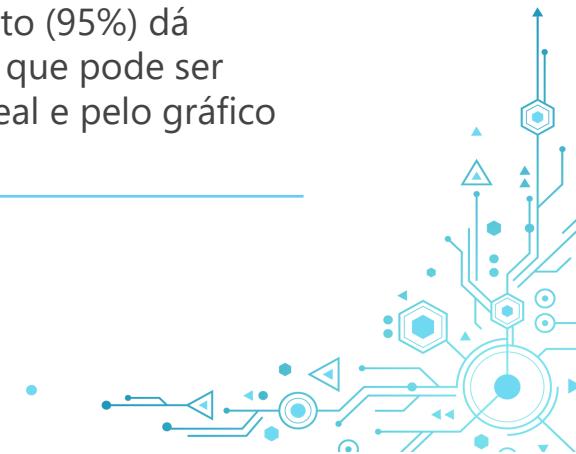
### Regressão linear c/ polinômio potência 2:

- R<sup>2</sup> ajustado: 0.9556366871746461
- Real | Previsto:
  - média: 65.09791 | 68.28762
  - min: 1 | 0
  - 25%: 1 | 2
  - 50%: 3 | 3
  - 75%: 11 | 9
  - max: 19822 | 25563
- sem previsões negativas



Indicadores com polinômio com potência 2 para base de treino	BASELINE	Regressão Linear ScikitLearn
VIÉS	4.31545	-3.18971
MSE	2087.44337	11642.50014
RMSE	45.68855	107.90042
MAE	4.41214	7.80948
MAPE	11.90706	27.98055

Com a **base treino**, a regressão linear teve **desempenho pior que a baseline**. O erro médio absoluto baixou para 7.8 casos (contra 4.4 casos da baseline). O R<sup>2</sup> ajustado alto (95%) dá mostras de overfitting com a base de treino, o que pode ser comprovado pela média prevista próxima da real e pelo gráfico de dispersão.



## 5.1.2. Método de Previsão Tradicional

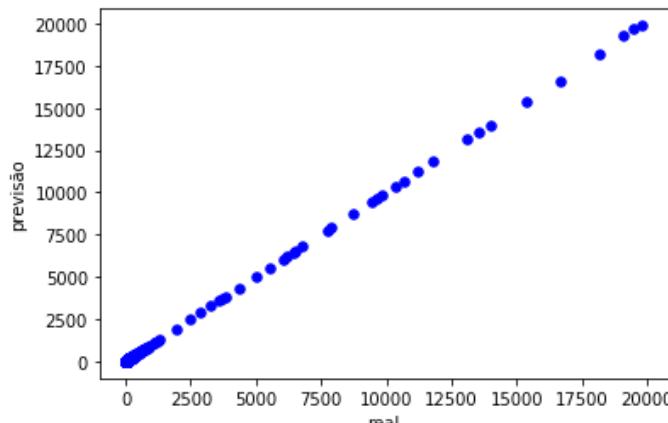
REGRESSÃO LINEAR | CASOS ACUMULADOS

91

BASE TREINO

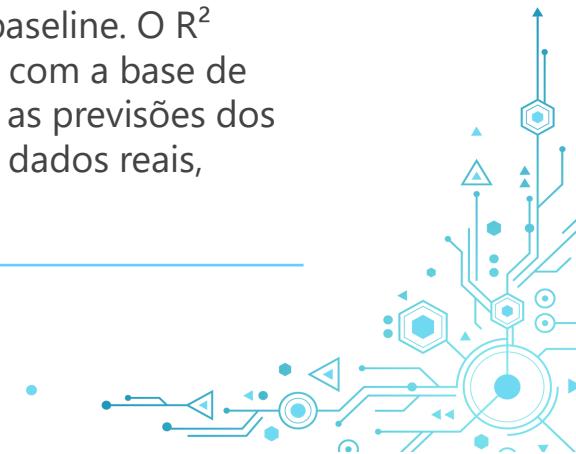
### Regressão linear c/ polinômio potência 3:

- R<sup>2</sup> ajustado: 1.008983128043202
- Real | Previsto
  - média: 65.09791 | 65.14625
  - min: 1 | 0
  - 25%: 1 | 1
  - 50%: 3 | 3
  - 75%: 11 | 11
  - max: 19822 | 19870
- sem previsões negativas



Indicadores com polinômio com potência 3 para base de treino	BASELINE	Regressão Linear ScikitLearn
VIÉS	4.31545	-0.04834
MSE	2087.44337	12.30554
RMSE	45.68855	3.50793
MAE	4.41214	0.71999
MAPE	11.90706	6.53939

Com a **base treino**, a regressão linear teve **desempenho melhor que a baseline pela primeira vez**: o erro médio absoluto baixou para 0.7 casos, contra 4.4 da baseline. O R<sup>2</sup> ajustado alto (100%) dá mostras de overfitting com a base de treino, que pode ser confirmado olhando para as previsões dos intervalos interquartis em comparação com os dados reais, pelas médias e pelo gráfico de dispersão.

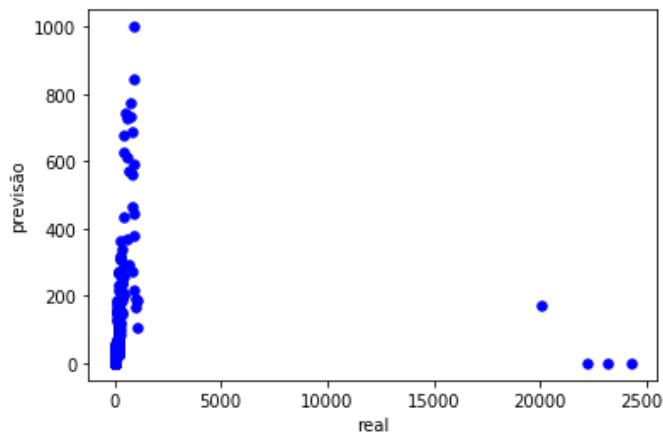


## 5.1.2. Método de Previsão Tradicional

REGRESSÃO LINEAR | CASOS ACUMULADOS

92

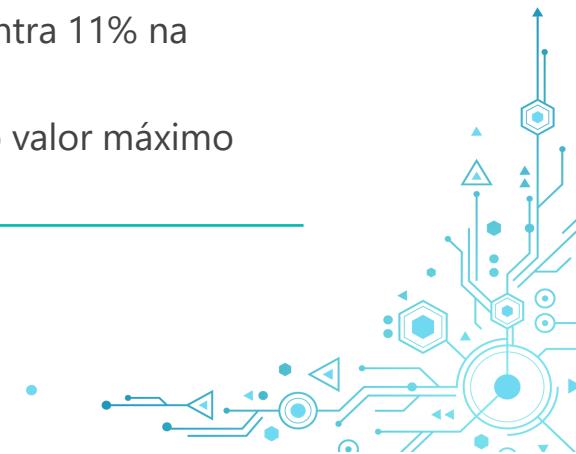
Regressão linear sem polinômio:
• Real   Previsto:
• média: 101.86723   26.79944
• min: 1   -1
• 25%: 2   2
• 50%: 5   6
• 75%: 19   16
• max: 24273   1000
• 2 previsões negativas (-1 previsto contra 22207 e 24273 real)



BASE TESTE		
Indicadores sem polinômio para base de teste	BASELINE	Regressão Linear ScikitLearn
VIÉS	5.69602	75.06778
MSE	4835.93082	1413445.71908
RMSE	69.54086	1188.88423
MAE	5.92383	82.12439
MAPE	11.07093	103.70583

Com a **base teste**, a regressão linear teve **desempenho pior que a baseline**. O erro médio absoluto foi de 82.1 casos, contra 5.9 casos na baseline (que apenas repete o resultado do dia anterior) e o MAPE aponta 103.7% de erro (contra 11% na baseline).

Analizando os intervalos interquartis, apenas o valor máximo previsto ficou muito discrepante do valor real.



## 5.1.2. Método de Previsão Tradicional

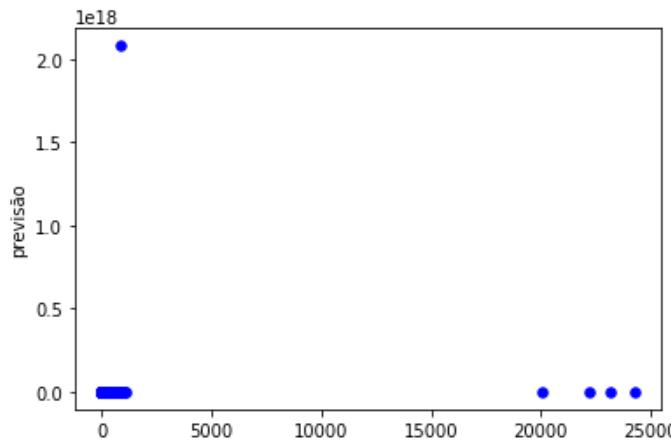
REGRESSÃO LINEAR | CASOS ACUMULADOS

93

BASE TESTE

### Regressão linear c/ polinômio potência 2:

- Real | Previsto:
  - média: 101.86723 | 1454674754612219.75
  - min: 1 | -1
  - 25%: 2 | 2
  - 50%: 5 | 4
  - 75%: 19 | 13
  - max: 24273 | 2081639545219988992
- 28 previsões negativas



@2020 LABDATA FIA. Copyright all rights reserved.

### Indicadores com polinômio com potência 2 para base de teste

	BASELINE	Regressão Linear ScikitLearn
VIÉS	5.69602	-1454674754612119.25
MSE	4835.93082	3028108452986498715868285903044608
RMSE	69.54086	55028251407676936
MAE	5.92383	1454674754612287
MAPE	11.07093	166058764005675.1875

Com a **base teste**, a regressão linear teve **desempenho pior que a baseline**. Todos os indicadores de erros explodiram, assim como a diferença entre o valor máximo previsto e real, invalidando o modelo.



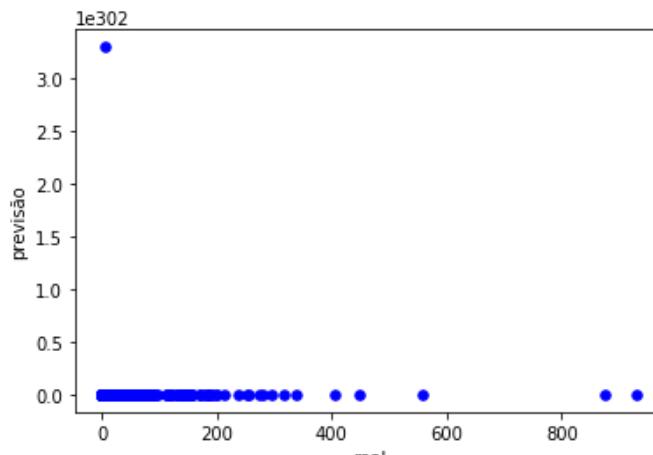
## 5.1.2. Método de Previsão Tradicional

REGRESSÃO LINEAR | CASOS ACUMULADOS

94

### BASE TESTE

Regressão linear c/ polinômio potência 3:
<ul style="list-style-type: none"><li>Real   Previsto:<ul style="list-style-type: none"><li>média: 101.86723   inf</li><li>min: 1   -1</li><li>25%: 2   -1</li><li>50%: 5   1</li><li>75%: 19   1033585816.5</li><li>max: 24273   inf</li></ul></li><li>398 previsões negativas</li></ul>



@2020 LABDATA FIA. Copyright all rights reserved.

Indicadores com polinômio com potência 3 para base de teste	BASELINE	Regressão Linear ScikitLearn
VIÉS	5.69602	Houve previsão infinita, fora da escala de medição
MSE	4835.93082	
RMSE	69.54086	
MAE	5.92383	
MAPE	11.07093	

Com a **base teste**, a regressão linear foi ainda **pior na comparação com a baseline**, com indicadores de erro fora de escala devido à previsão infinita no valor máximo, invalidando o modelo.



## 5.1.2. Método de Previsão Tradicional

REGRESSÃO LINEAR | CASOS ACUMULADOS

95

### Conclusões:

- o método de previsão tradicional com regressão linear do Scikit Learn teve desempenho pior que a baseline para prever casos acumulados.
- a aplicação do polinômio melhora a performance na base treino, mas estraga a previsão na base teste, principalmente nas previsões acima do 3º quartil.

Vamos testar outro método de regressão linear com a biblioteca do StatsModels.

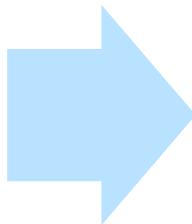


## 5.1.2. Método de Previsão Tradicional

REGRESSÃO LINEAR | CASOS ACUMULADOS

96

Com acesso as estatísticas completas da regressão linear, eliminamos todas as variáveis com p-valor > 0.05, ficando com o seguinte conjunto de explicativas:



'mortes\_acumuladas\_menos6d'  
'mortes\_acumuladas\_menos14d'  
'casos\_acumulados\_menos1d'  
'casos\_acumulados\_menos13d'  
'capital\_S'  
'Hospital/DIA\_SUS'  
'papel\_Metropolitana'  
'papel\_Interior'  
'dias\_epidemiológicos'  
'Obstétrico\_SUS'  
'Obstétrico\_Não\_SUS'  
'Hospital/DIA\_Não\_SUS'  
'Clínicos\_SUS'  
'Cirúrgicos\_SUS'



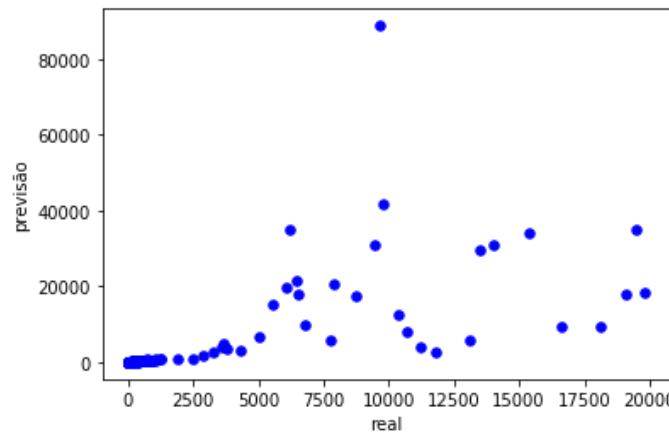
## 5.1.2. Método de Previsão Tradicional

REGRESSÃO LINEAR | CASOS ACUMULADOS

97

BASE TREINO

Regressão linear sem polinômio:	
• R <sup>2</sup> ajustado: 0.772	
• Real   Previsto:	
• média: 65.09791   93.06192	
• min: 1   -1	
• 25%: 1   2	
• 50%: 3   4	
• 75%: 11   11	
• max: 19822   88797	
• 4 previsões negativas (-1 previsto contra 1 real)	



Indicadores sem polinômio para base de treino	Baseline	Regressão Linear ScikitLearn	Regressão Linear StatsModels OLS
VIÉS	4.31545	-14.84737	-27.96401
MSE	2087.44337	386216.65318	1482975.94555
RMSE	45.68855	621.46332	1217.77500
MAE	4.41214	40.32319	58.86027
MAPE	11.90706	95.74353	96.80734

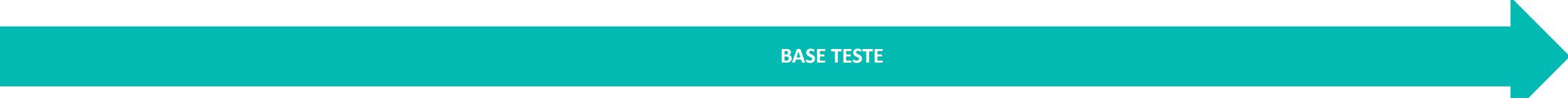
Mesmo eliminando variáveis com p-valor > 0.05, a **regressão linear OLS do StatsModels teve performance pior** do que a regressão do scikitlearn com todas as variáveis.



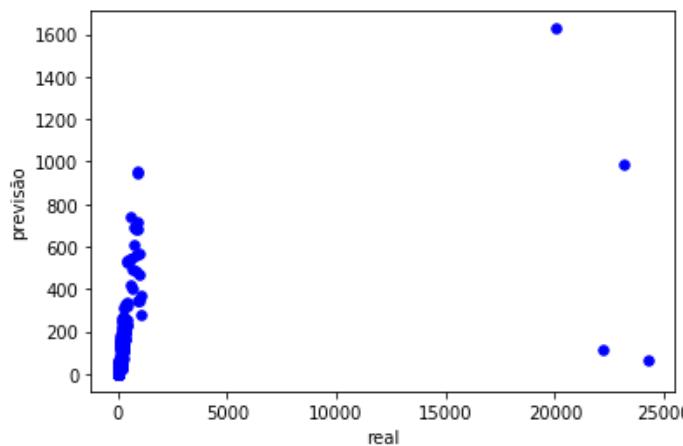
## 5.1.2. Método de Previsão Tradicional

REGRESSÃO LINEAR | CASOS ACUMULADOS

98



Regressão linear sem polinômio:
• Real   Previsto:
• média: 101.86723   29.74004
• min: 1   0
• 25%: 2   2
• 50%: 5   6
• 75%: 19   17
• max: 24273   1627
• sem previsões negativas



@2020 LABDATA FIA. Copyright all rights reserved.

Indicadores sem polinômio para base de teste	Baseline	Regressão Linear ScikitLearn	Regressão Linear StatsModels OLS
VIÉS	5.69602	75.06778	72.12718
MSE	4835.93082	1413445.71908	1335712.03075
RMSE	69.54086	1188.88423	1155.73009
MAE	5.92383	82.12439	78.14256
MAPE	11.07093	103.70583	104.45193

Com a **base teste**, a regressão linear OLS do StatsModels teve **desempenho pior que a baseline** e um **pouco melhor que a regressão do SciKitLearn** nos indicadores de erro médio absoluto, viés e RMSE, embora com leve piora no MAPE.

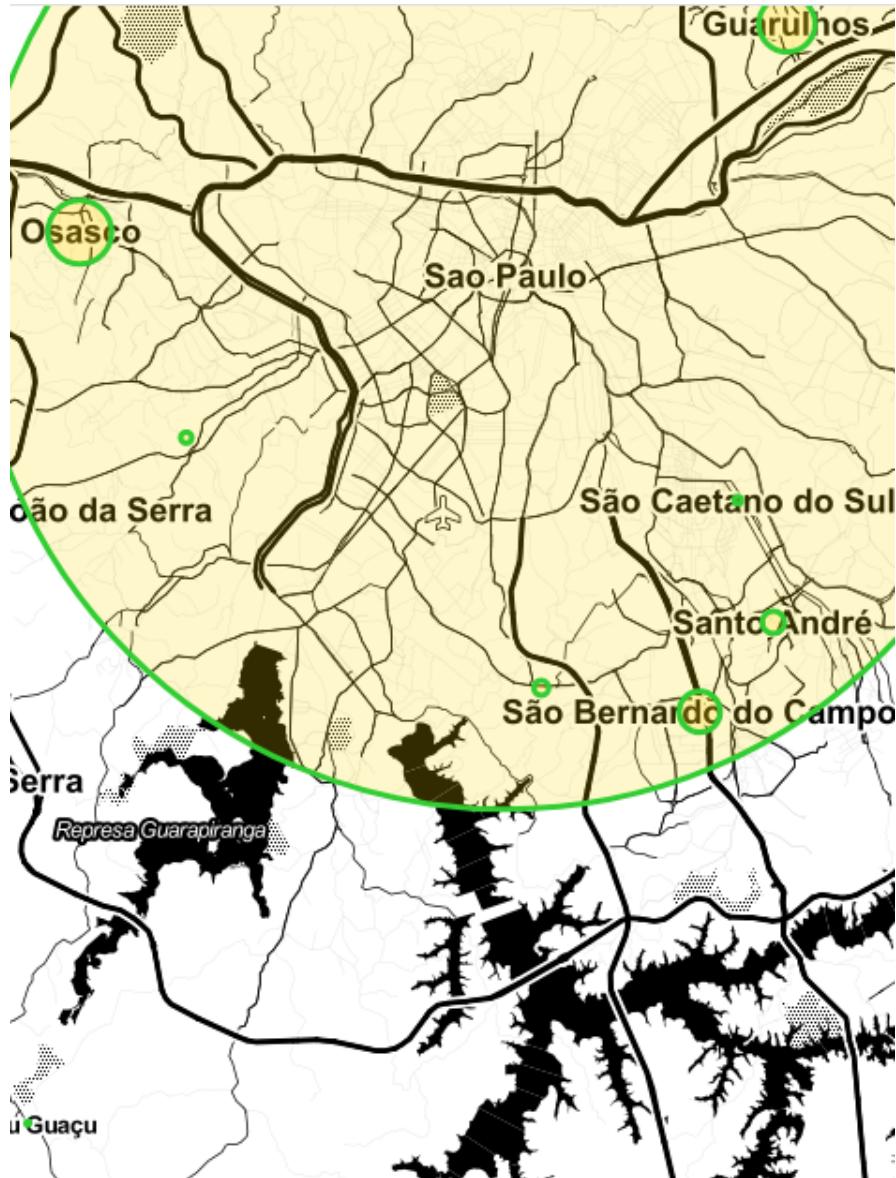
Analizando os intervalos interquartis, apenas o valor máximo previsto ficou muito discrepante do valor real.



## 5.1.2. Método de Previsão Tradicional

REGRESSÃO LINEAR | MORTES ACUMULADAS

99



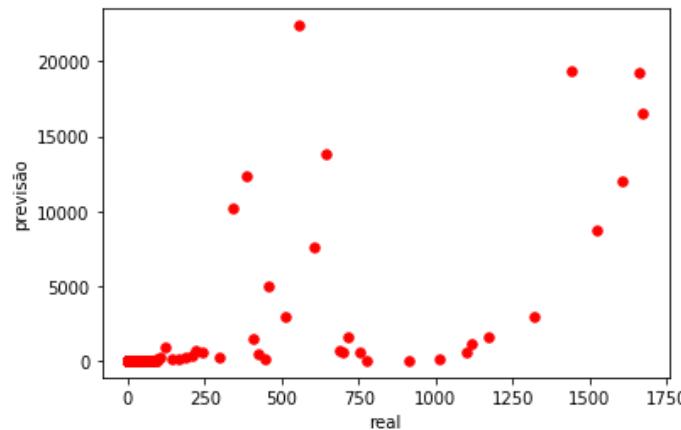
## 5.1.2. Método de Previsão Tradicional

REGRESSÃO LINEAR | MORTES ACUMULADAS

100

BASE TREINO

Regressão linear sem polinômio:	
• R <sup>2</sup> ajustado:	0.38321765694473686
• Real   Previsto:	
• média:	4.98669   22.92260
• min:	0   0
• 25%:	0   0
• 50%:	0   0
• 75%:	1   1
• max:	1673   22398
• sem previsões negativas	



Indicadores sem polinômio para base de treino	BASELINE	Regressão Linear ScikitLearn
VIÉS	0.35646	-17.93590
MSE	16.43061	269773.48832
RMSE	4.05347	519.39724
MAE	0.36353	21.56817
MAPE	9.64655	inf

Com a **base treino**, a regressão linear teve **desempenho pior que a baseline**. O erro médio absoluto foi de 21.5 mortes, contra 0.3 mortes na baseline (que apenas repete o resultado do dia anterior) e o MAPE aponta erro infinito (contra 9.6% na baseline).

Analizando os intervalos interquartis, apenas o valor máximo previsto ficou muito discrepante do valor real.



## 5.1.2. Método de Previsão Tradicional

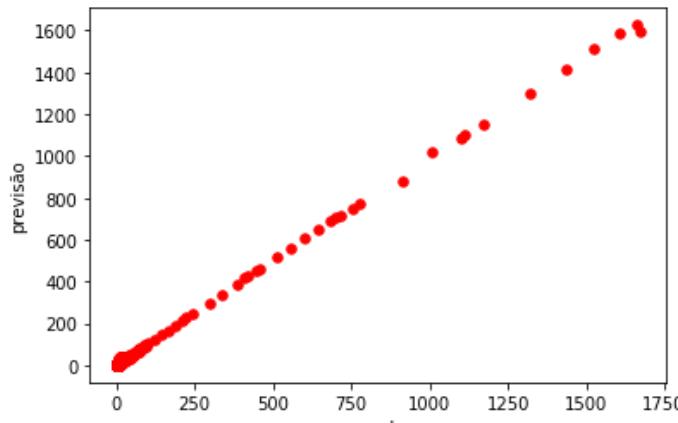
REGRESSÃO LINEAR | MORTES ACUMULADAS

101

BASE TREINO

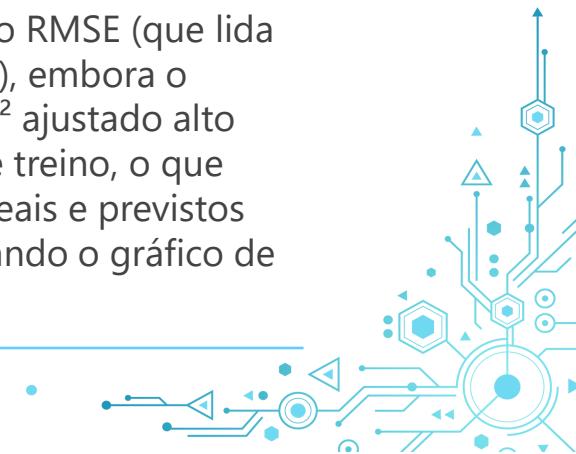
### Regressão linear c/ polinômio potência 2:

- $R^2$  ajustado: 0.9385217982793314
- Real | Previsto:
  - média: 4.98669 | 4.99919
  - min: 0 | 0
  - 25%: 0 | 0
  - 50%: 0 | 0
  - 75%: 1 | 1
  - max: 1673 | 1626
- sem previsões negativas



Indicadores com polinômio com potência 2 para base de treino	BASELINE	Regressão Linear ScikitLearn
VIÉS	0.35646	-0.01249
MSE	16.43061	4.00679
RMSE	4.05347	2.00170
MAE	0.36353	0.45220
MAPE	9.64655	inf

Com a **base treino**, a regressão linear teve **desempenho semelhante à baseline**. O erro médio absoluto ainda ficou um pouco acima da baseline (0.4 contra 0.3), mas o RMSE (que lida melhor com outliers) esteve abaixo (2 contra 4), embora o MAPE continue apontando para o infinito. O  $R^2$  ajustado alto (93%) dá mostras de overfitting com a base de treino, o que pode ser confirmado comparando os valores reais e previstos dos intervalos interquartis, da média e observando o gráfico de dispersão.



## 5.1.2. Método de Previsão Tradicional

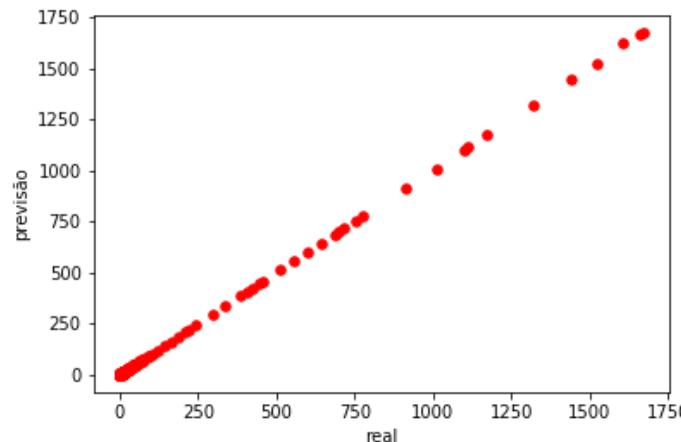
REGRESSÃO LINEAR | MORTES ACUMULADAS

102

BASE TREINO

### Regressão linear c/ polinômio potência 3:

- R<sup>2</sup> ajustado: 0.9526883460758565
- Real | Previsto
  - média: 4.98669 | 4.97637
  - min: 0 | 0
  - 25%: 0 | 0
  - 50%: 0 | 0
  - 75%: 1 | 11
  - max: 1673 | 1673
- sem previsões negativas



Indicadores com polinômio com potência 3 para base de treino	BASELINE	Regressão Linear ScikitLearn
VIÉS	0.35646	0.01032
MSE	16.43061	0.08365
RMSE	4.05347	0.28922
MAE	0.36353	0.03286
MAPE	9.64655	4.80707

Com a **base treino**, a regressão linear teve **desempenho melhor que a baseline pela primeira vez**: o erro médio absoluto baixou para 0.03 mortes, e todos os outros indicadores de erros também conseguiram superar a baseline. O R<sup>2</sup> ajustado alto (95%) dá mostras de overfitting com a base de treino, que pode ser confirmado olhando para as previsões dos intervalos interquartis em comparação com os dados reais, para as médias e para o gráfico de dispersão.



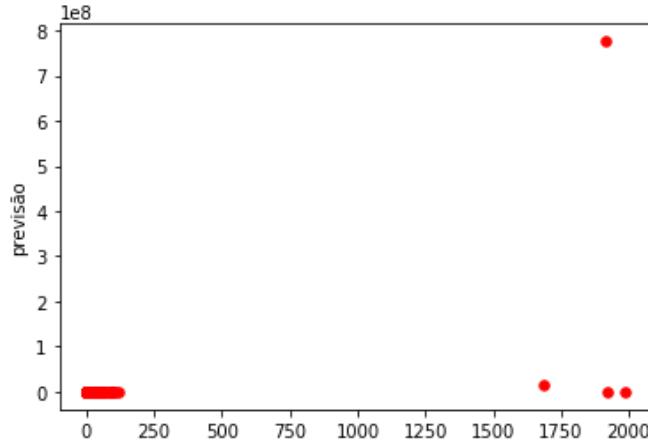
## 5.1.2. Método de Previsão Tradicional

REGRESSÃO LINEAR | MORTES ACUMULADAS

103



Regressão linear sem polinômio:
• Real   Previsto:
• média: 8.34731   552570.24808
• min: 0   -1
• 25%: 0   0
• 50%: 0   0
• 75%: 2   1
• max: 1986   776014980
• 2 previsões negativas (-1 previsto contra 1919 e 1986 real)

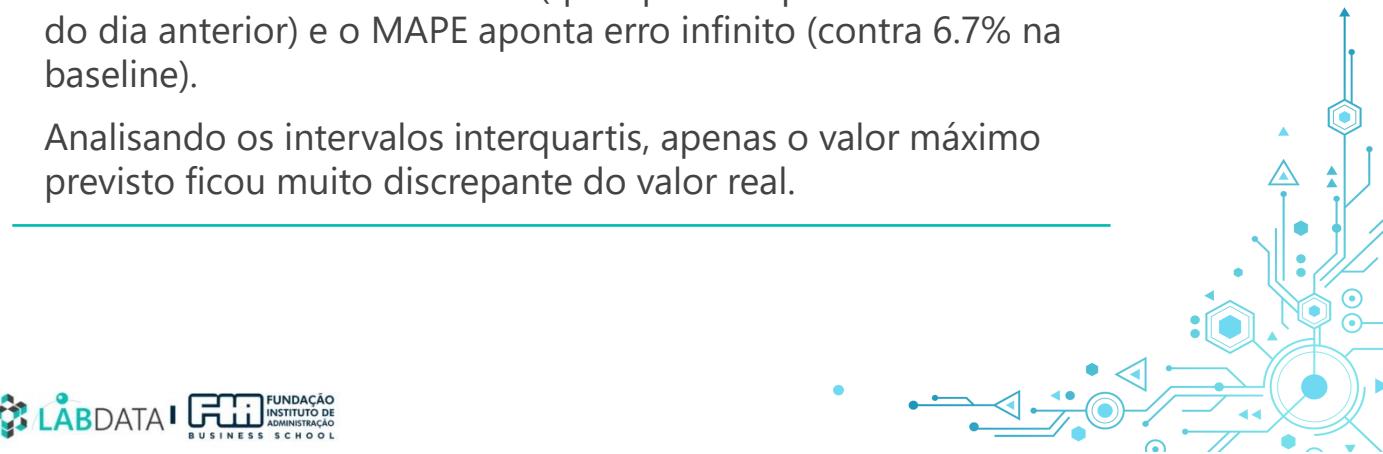


@2020 LABDATA FIA. Copyright all rights reserved.

Indicadores sem polinômio para base de teste	BASELINE	Regressão Linear ScikitLearn
VIÉS	0.40601	-552561.90077
MSE	44.75681	420973229199271.18750
RMSE	6.69005	20517632.15382
MAE	0.47729	552572.77428
MAPE	6.72089	inf

Com a **base teste**, a regressão linear teve **desempenho pior que a baseline**. O erro médio absoluto foi de 552572 mortes, contra 0.4 mortes na baseline (que apenas repete o resultado do dia anterior) e o MAPE aponta erro infinito (contra 6.7% na baseline).

Analizando os intervalos interquartis, apenas o valor máximo previsto ficou muito discrepante do valor real.



## 5.1.2. Método de Previsão Tradicional

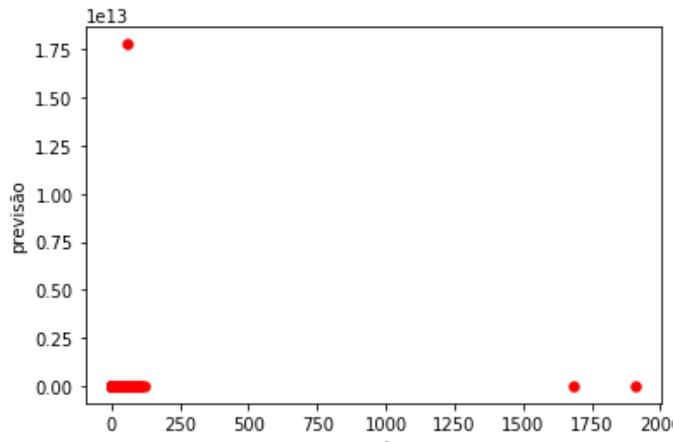
REGRESSÃO LINEAR | MORTES ACUMULADAS

104

BASE TESTE

### Regressão linear c/ polinômio potência 2:

- Real | Previsto:
  - média: 8.34731 | inf
  - min: 0 | -1
  - 25%: 0 | 0
  - 50%: 0 | 0
  - 75%: 2 | 1
  - max: 1986 | inf
- 16 previsões negativas (-1 previsto contra variação entre 23 e 1910)



Indicadores com polinômio com potência 2 para base de teste	BASELINE	Regressão Linear ScikitLearn
VIÉS	0.40601	Houve previsão infinita, fora da escala de medição
MSE	44.75681	
RMSE	6.69005	
MAE	0.47729	
MAPE	6.72089	

Com a **base teste**, a regressão linear foi ainda **pior na comparação com a baseline**, com indicadores de erro fora de escala devido à previsão infinita no valor máximo, invalidando o modelo.



## 5.1.2. Método de Previsão Tradicional

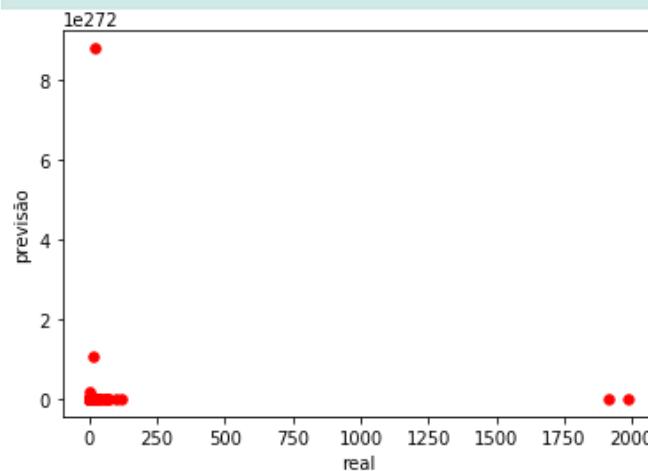
REGRESSÃO LINEAR | MORTES ACUMULADAS

105

### BASE TESTE

#### Regressão linear c/ polinômio potência 3:

- Real | Previsto:
  - média: 8.34731 | inf
  - min: 0 | -1
  - 25%: 0 | 0
  - 50%: 0 | 0
  - 75%: 2 | 1
  - max: 1986 | inf
- 238 previsões negativas



Indicadores com polinômio com potência 3 para base de teste	BASELINE	Regressão Linear ScikitLearn
VIÉS	0.40601	Houve previsão infinita, fora da escala de medição
MSE	44.75681	
RMSE	6.69005	
MAE	0.47729	
MAPE	6.72089	

Com a **base teste**, a regressão linear foi ainda **pior na comparação com a baseline**, com indicadores de erro fora de escala devido à previsão infinita no valor máximo, invalidando o modelo.



## 5.1.2. Método de Previsão Tradicional

REGRESSÃO LINEAR | MORTES ACUMULADAS

106

### Conclusões:

- o método de previsão tradicional com regressão linear do Scikit Learn teve desempenho pior que a baseline para prever mortes acumuladas.
- a aplicação do polinômio melhora a performance na base treino, a ponto de dar overfitting, mas estraga a previsão na base teste, principalmente nas previsões acima do 3º quartil.

Vamos testar outro método de regressão linear com a biblioteca do StatsModels.

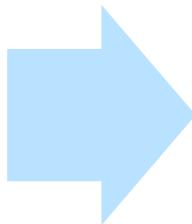


## 5.1.2. Método de Previsão Tradicional

REGRESSÃO LINEAR | MORTES ACUMULADAS

107

Com acesso as estatísticas completas da regressão linear, eliminamos todas as variáveis com p-valor > 0.05, ficando com o seguinte conjunto de explicativas:



'mortes\_acumuladas\_menos1d'  
'mortes\_acumuladas\_menos2d'  
'mortes\_acumuladas\_menos14d'  
'mortes\_acumuladas\_menos13d'  
'mortes\_acumuladas\_menos12d'  
'mortes\_acumuladas\_menos11d'  
'casos\_acumulados\_menos7d'  
'casos\_acumulados\_menos3d'  
'casos\_acumulados\_menos1d'  
'casos\_acumulados\_menos13d'  
'casos\_acumulados\_menos10d'  
'capital\_S', 'Hospital/DIA\_SUS'



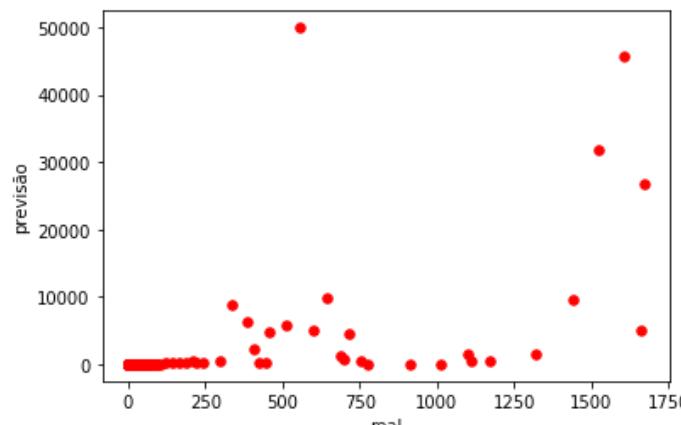
## 5.1.2. Método de Previsão Tradicional

REGRESSÃO LINEAR | MORTES ACUMULADAS

108

BASE TREINO

Regressão linear sem polinômio:	
• R <sup>2</sup> ajustado: 0.379	
• Real   Previsto:	
• média: 4.98669   31.12873	
• min: 0   0	
• 25%: 0   0	
• 50%: 0   0	
• 75%: 1   1	
• max: 1673   49936	
• sem previsões negativas	



Indicadores sem polinômio para base de treino	Baseline	Regressão Linear ScikitLearn	Regressão Linear StatsModels OLS
VIÉS	0.35646	-17.93590	-26.14204
MSE	16.43061	269773.48832	854682.35741
RMSE	4.05347	519.39724	924.49032
MAE	0.36353	21.56817	29.96985
MAPE	9.64655	inf	inf

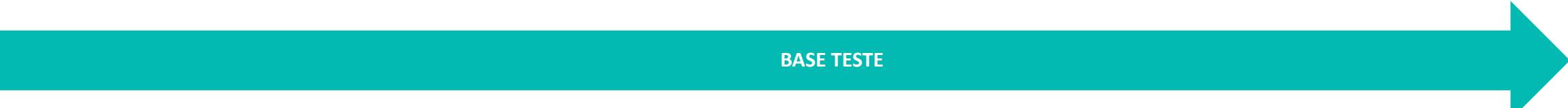
Mesmo eliminando variáveis com p-valor > 0.05, a **regressão linear OLS do StatsModels teve performance pior** do que a regressão do scikitlearn com todas as variáveis.



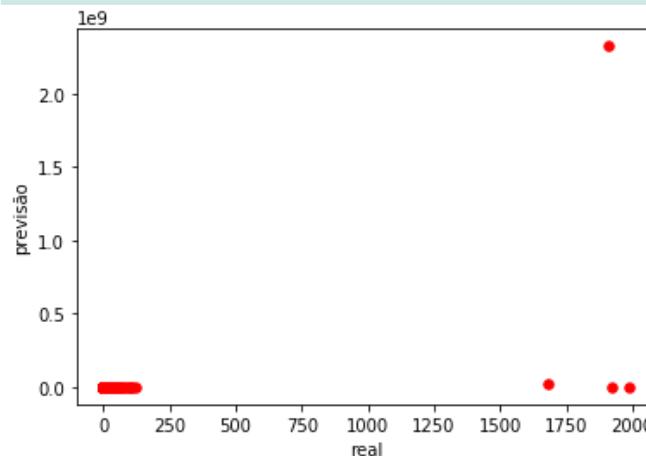
## 5.1.2. Método de Previsão Tradicional

REGRESSÃO LINEAR | MORTES ACUMULADAS

109



Regressão linear sem polinômio:
• Real   Previsto:
• média: 8.34731   1638747.46331
• min: 0   -1
• 25%: 0   0
• 50%: 0   0
• 75%: 2   1
• max: 1986   2326533822
• 1 previsão negativa (-1 previsto contra 1986 real)



Indicadores sem polinômio para base de teste	Baseline	Regressão Linear ScikitLearn	Regressão Linear StatsModels OLS
VIÉS	0.40601	-552561.90077	-1638739.11600
MSE	44.75681	420973229199271.18750	3782734726619824.5
RMSE	6.69005	20517632.15382	61503940.74057
MAE	0.47729	552572.77428	1638749.25157
MAPE	6.72089	inf	inf

Com a **base teste**, a regressão linear OLS do StatsModels teve **desempenho pior que a baseline e muito pior que a regressão do SciKitLearn** em todos os indicadores de erro.

Analizando os intervalos interquartis, apenas o valor máximo previsto ficou muito discrepante do valor real.



## 5.2. Modelagem com Estatística Tradicional

APLICAÇÃO DE MODELO | ÁRVORE DE DECISÃO

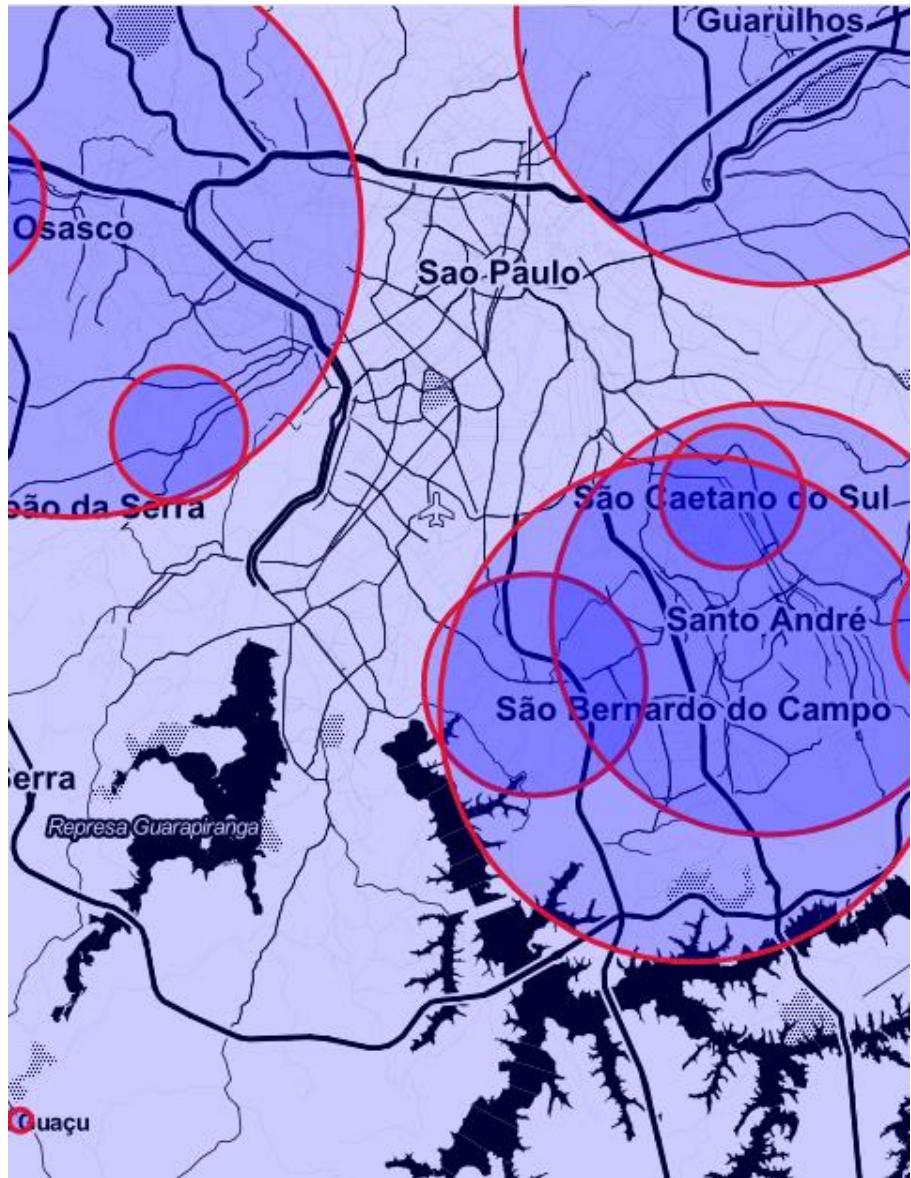
110

# ÁRVORE DE DECISÃO

## 5.2.1. Método de Previsão em Laço

ÁRVORE DE DECISÃO | CASOS ACUMULADOS

111



@2020 LABDATA FIA. Copyright all rights reserved.

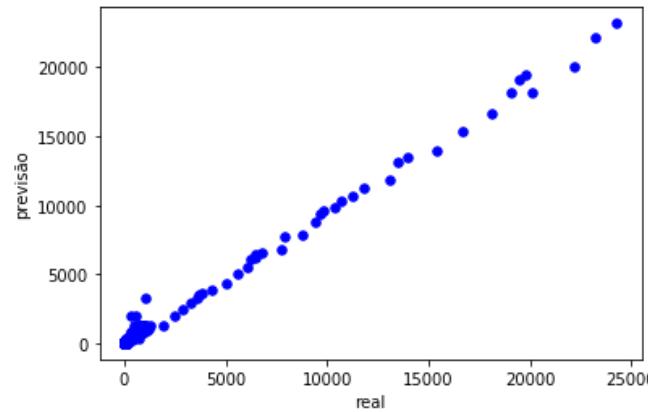
## 5.2.1. Método de Previsão em Laço

ÁRVORE DE DECISÃO | CASOS ACUMULADOS

112

### BASE TESTE – SEM MODIFICAÇÃO DA DISTRIBUIÇÃO DA TARGET

Árvore de decisão sem modificação da distribuição target:
• Real   Previsto:
• Média: 71.08847   69.82943
• min: 1   1
• 25%: 1   1
• 50%: 3   4
• 75%: 13   14
• max: 24273   23187
• sem previsões negativas



Indicadores para base de teste	BASELINE	Árvore de Decisão (laço)
VIÉS	4.54008	1.25904
MSE	2534.639	4375.49875
RMSE	50.3452	66.14755
MAE	4.6581	6.74528
MAPE	11.77102	18.34111

Com árvore de decisão sem modificação da distribuição da target, o laço funcionou durante todos os **128 dias**, com indicadores de erros aumentando gradativamente.

Com a **base teste**, a árvore de decisão teve **desempenho pouco pior que a baseline, mas muito melhor do que a regressão linear**. O erro médio absoluto foi de 6.7 casos, contra 4.6 casos na baseline (que apenas repete o resultado do dia anterior).

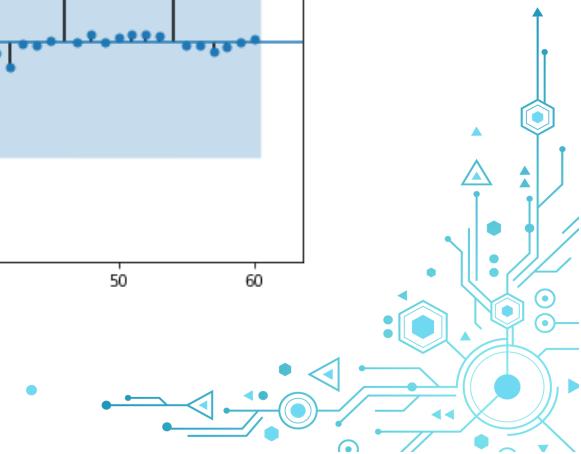
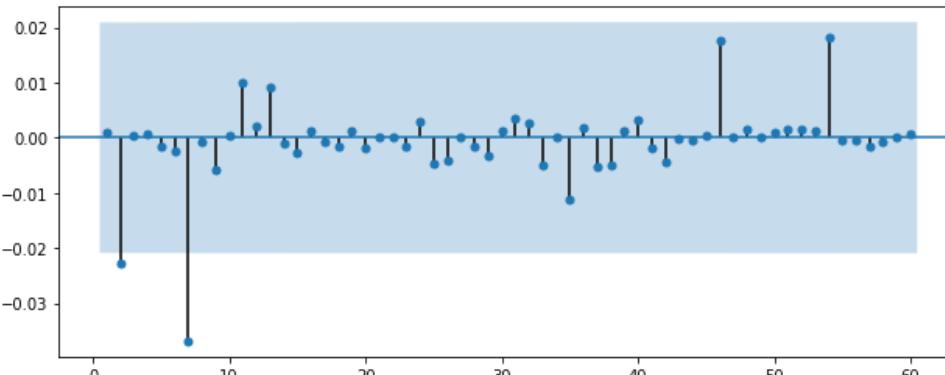
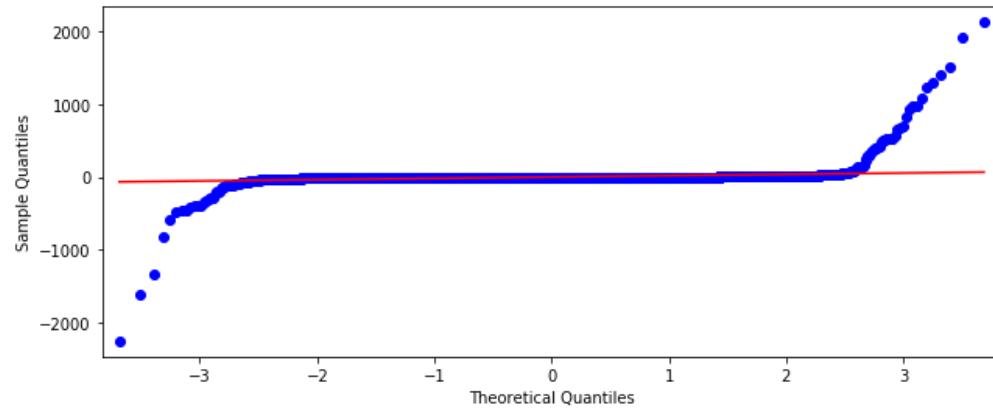
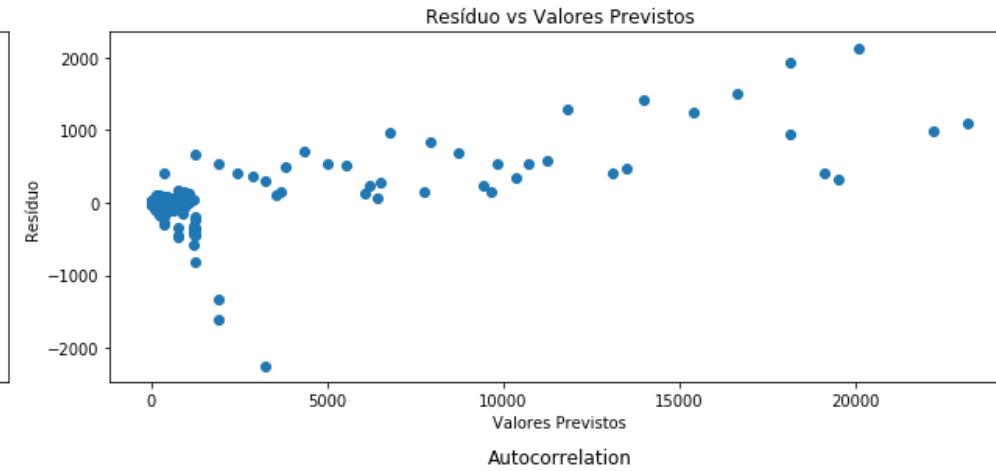
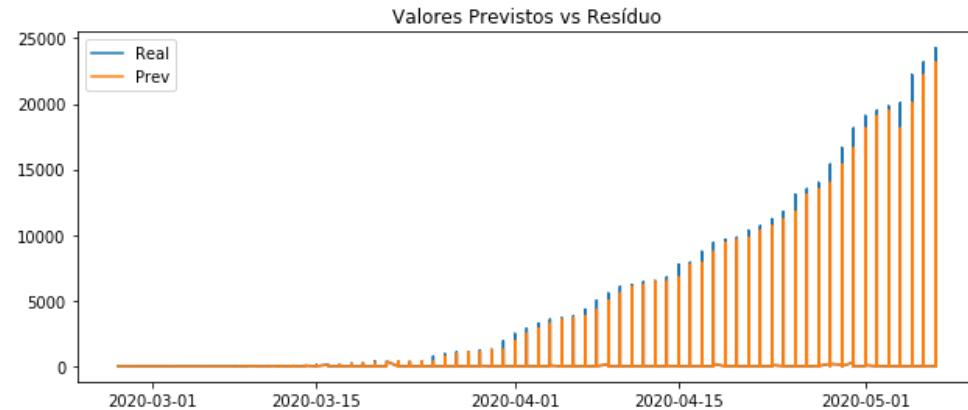


## 5.2.1. Método de Previsão em Laço

ÁRVORE DE DECISÃO | CASOS ACUMULADOS

113

BASE TESTE – SEM MODIFICAÇÃO DA DISTRIBUIÇÃO DA TARGET



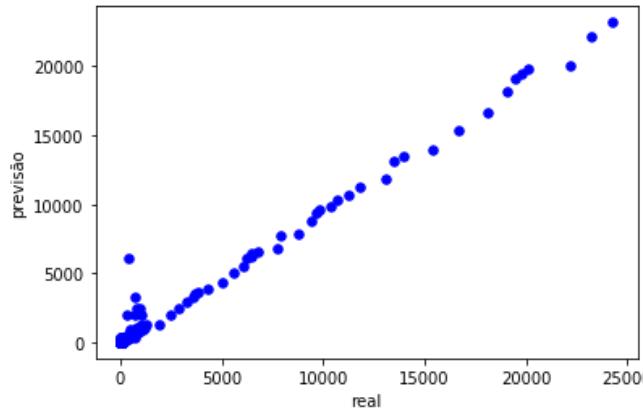
## 5.2.1. Método de Previsão em Laço

ÁRVORE DE DECISÃO | CASOS ACUMULADOS

114

### BASE TESTE – COM MODIFICAÇÃO DA DISTRIBUIÇÃO DA TARGET

Árvore de decisão com modificação da distribuição da target:
• Real   Previsto:
• Média: 71.08847   70.71117
• min: 1   1
• 25%: 1   1
• 50%: 3   3
• 75%: 13   13
• max: 24273   23187
• sem previsões negativas



Indicadores para base de teste	BASELINE	Árvore de Decisão (laço)
VIÉS	4.54008	0.37730
MSE	2534.639	8188.02024
RMSE	50.3452	90.48768
MAE	4.6581	7.48761
MAPE	11.77102	18.96010

Com árvore de decisão com modificação da distribuição da target, o laço funcionou durante todos os **128 dias**, com indicadores de erros aumentando gradativamente. Com a **base teste**, a árvore de decisão teve **desempenho pouco pior que a baseline, mas muito melhor do que a regressão linear**. O erro médio absoluto foi de 7.4 casos, contra 4.6 casos na baseline. Na comparação com a árvore sem modificação da target, a média das previsões ficou mais próxima da média real, o viés diminuiu, mas o erro absoluto aumentou, devido às previsões do início do gráfico de dispersão.

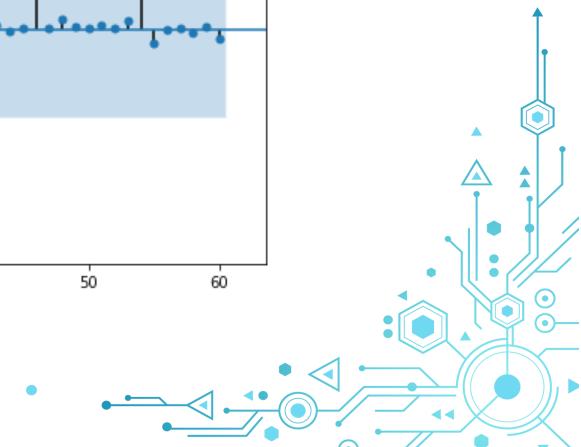
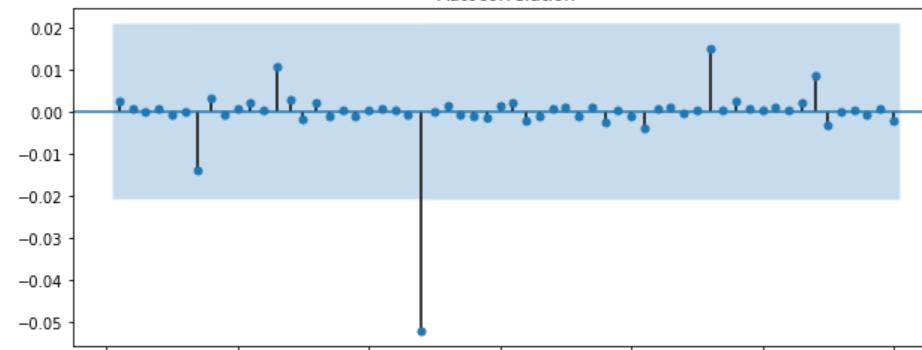
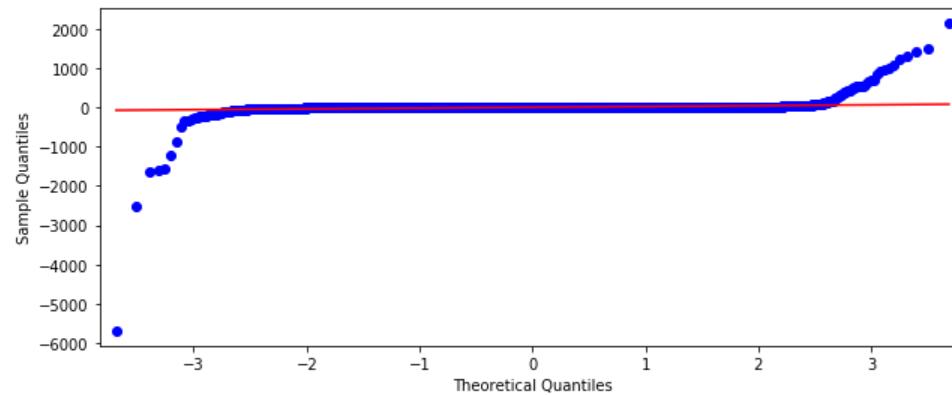
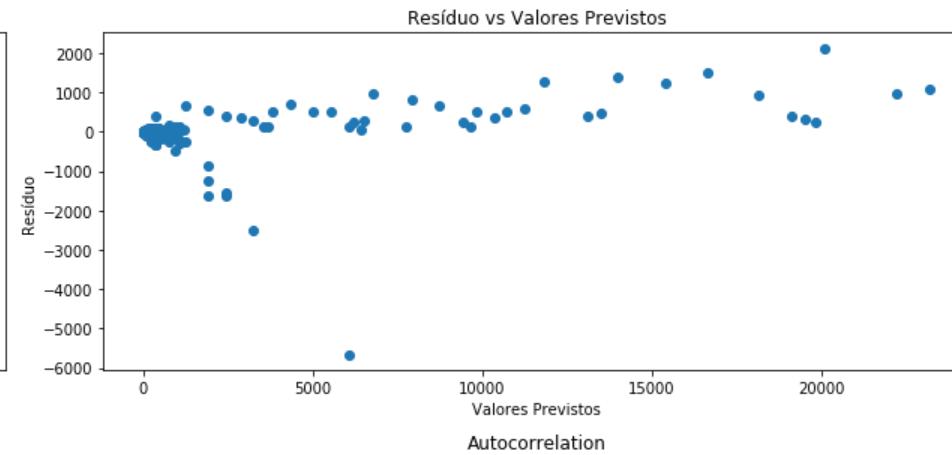
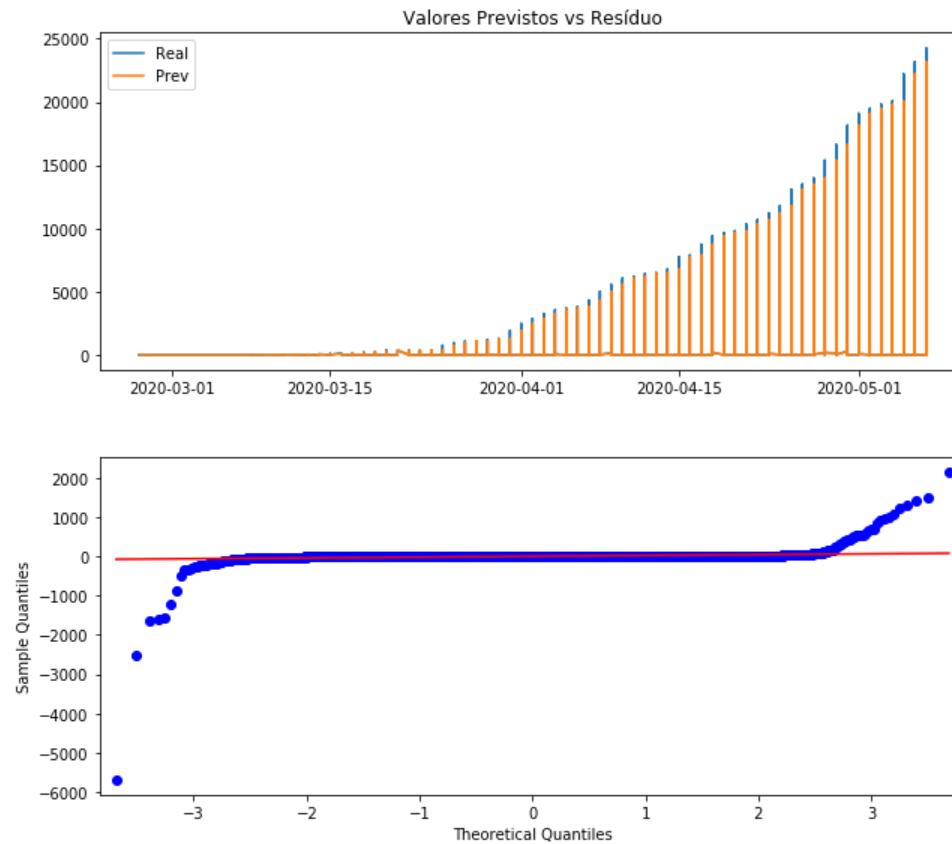


## 5.2.1. Método de Previsão em Laço

ÁRVORE DE DECISÃO | CASOS ACUMULADOS

115

BASE TESTE – COM MODIFICAÇÃO DA DISTRIBUIÇÃO DA TARGET



## 5.2.1. Método de Previsão em Laço

ÁRVORE DE DECISÃO | CASOS ACUMULADOS

116

### Conclusões:

- o método de previsão em laço com árvore de decisão teve performance muito superior à regressão linear para prever casos acumulados.
- ainda assim, perdeu por pouco para a baseline.



## 5.2.1. Método de Previsão em Laço

ÁRVORE DE DECISÃO | MORTES ACUMULADAS



117

PREVISÃO  
EM LAÇO

ÁRVORE DE  
DECISÃO

MORTES  
ACUMULADAS



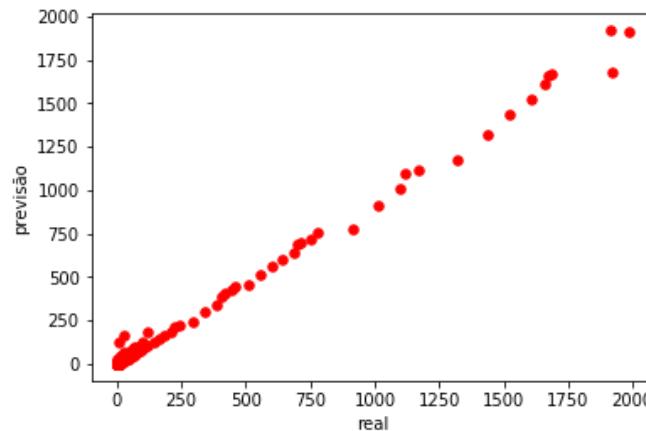
## 5.2.1. Método de Previsão em Laço

ÁRVORE DE DECISÃO | MORTES ACUMULADAS

118

### BASE TESTE – SEM MODIFICAÇÃO DA DISTRIBUIÇÃO DA TARGET

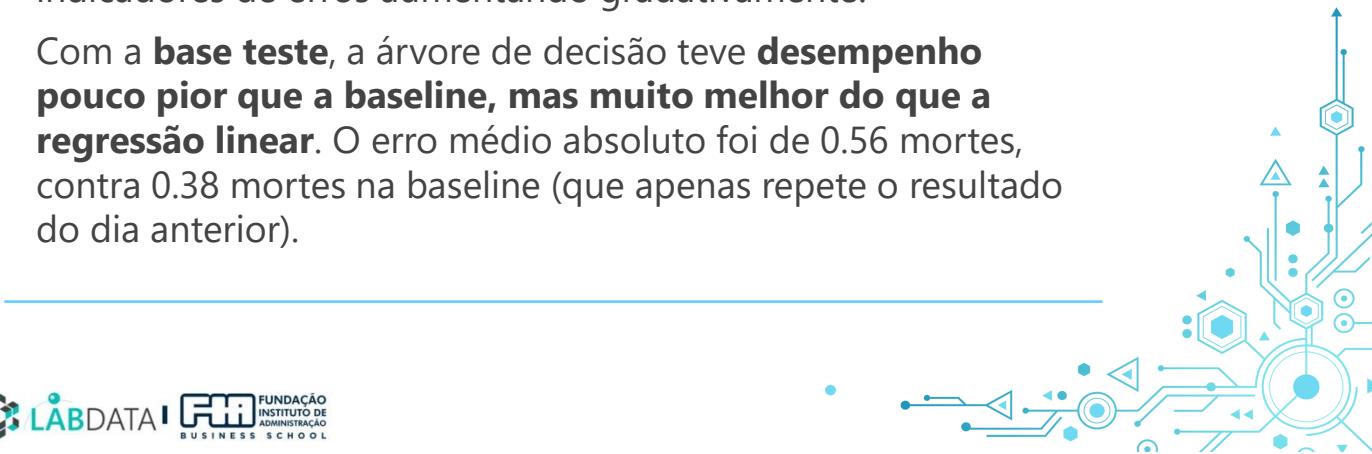
Árvore de decisão sem modificação da distribuição target:
• Real   Previsto:
• Média: 5.53411   5.40675
• min: 0   -9223372036854775808
• 25%: 0   0
• 50%: 0   0
• 75%: 1   1
• max: 1986   1919
• sem previsões negativas



Indicadores para base de teste	BASELINE	Árvore de Decisão (laço)
VIÉS	0.36453	0.12736
MSE	21.03945	26.14078
RMSE	4.58688	5.11281
MAE	0.38204	0.56811
MAPE	nan	inf

Com árvore de decisão sem modificação da distribuição da target, o laço funcionou durante todos os **128 dias**, com indicadores de erros aumentando gradativamente.

Com a **base teste**, a árvore de decisão teve **desempenho pouco pior que a baseline, mas muito melhor do que a regressão linear**. O erro médio absoluto foi de 0.56 mortes, contra 0.38 mortes na baseline (que apenas repete o resultado do dia anterior).

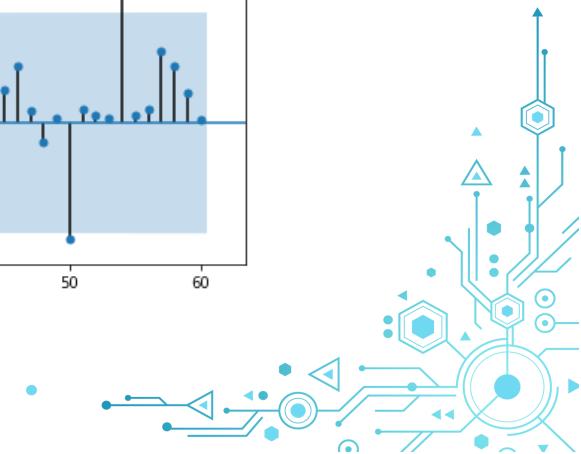
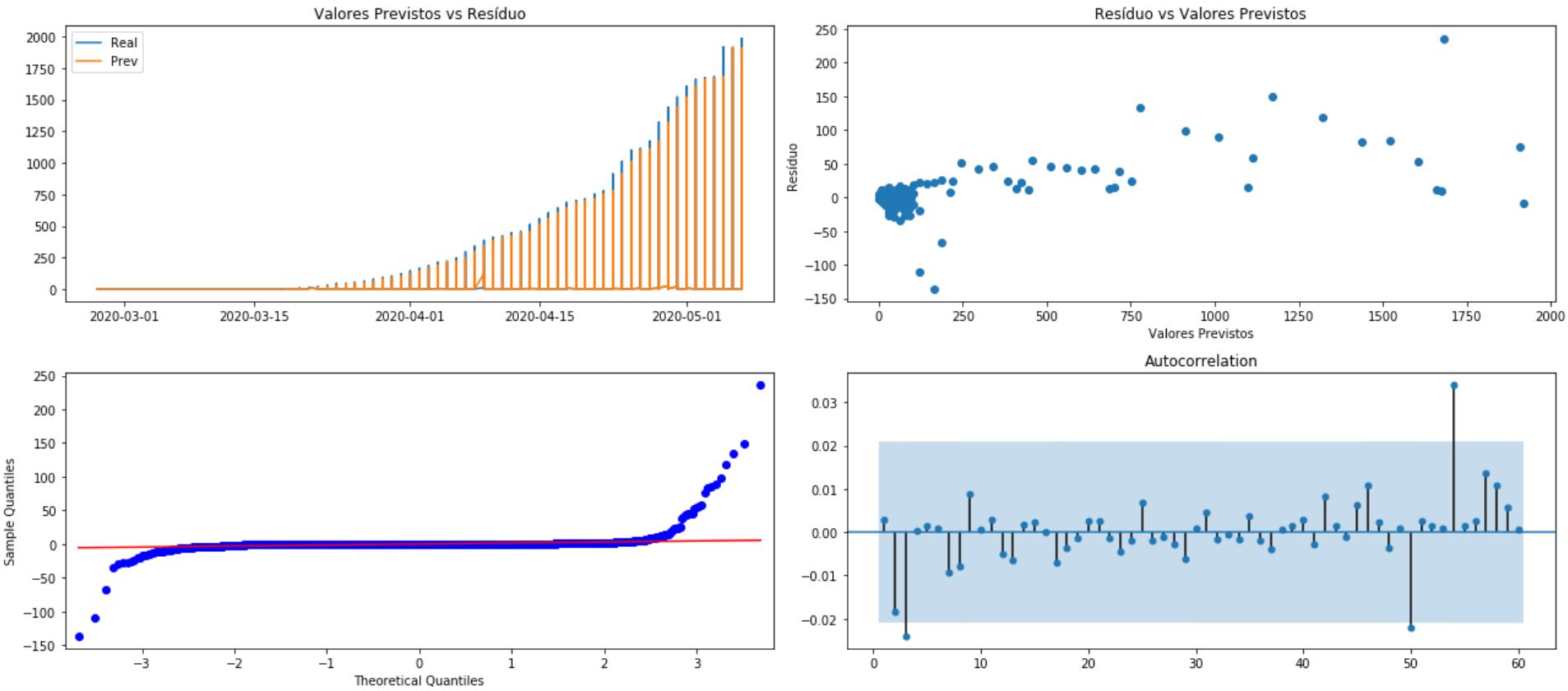


## 5.2.1. Método de Previsão em Laço

ÁRVORE DE DECISÃO | MORTES ACUMULADAS

119

BASE TESTE – SEM MODIFICAÇÃO DA DISTRIBUIÇÃO DA TARGET



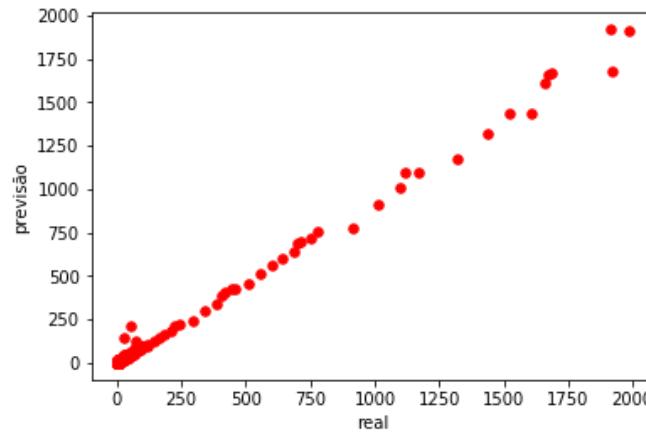
## 5.2.1. Método de Previsão em Laço

ÁRVORE DE DECISÃO | MORTES ACUMULADAS

120

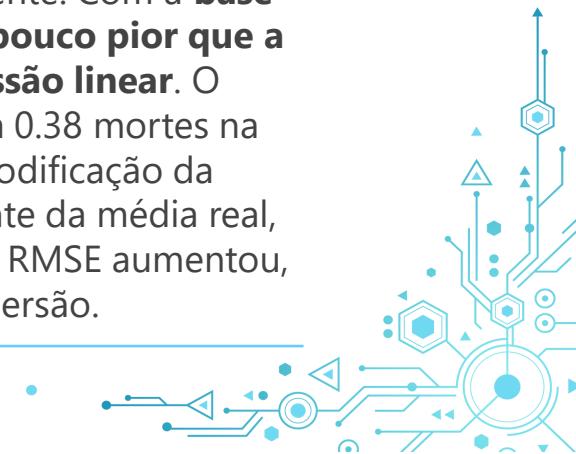
### BASE TESTE – COM MODIFICAÇÃO DA DISTRIBUIÇÃO DA TARGET

Árvore de decisão com modificação da distribuição target:
• Real   Previsto:
• Média: 5.53411   5.37901
• min: 0   -9223372036854775808
• 25%: 0   0
• 50%: 0   0
• 75%: 1   1
• max: 1986   1919
• sem previsões negativas



Indicadores para base de teste	BASELINE	Árvore de Decisão (laço)
VIÉS	0.36453	0.15511
MSE	21.03945	29.16375
RMSE	4.58688	5.40035
MAE	0.38204	0.56561
MAPE	nan	inf

Com árvore de decisão com modificação da distribuição da target, o laço funcionou durante todos os **128 dias**, com indicadores de erros aumentando gradativamente. Com a **base teste**, a árvore de decisão teve **desempenho pouco pior que a baseline, mas muito melhor do que a regressão linear**. O erro médio absoluto foi de 0.56 mortes, contra 0.38 mortes na baseline. Na comparação com a árvore sem modificação da target, a média das previsões ficou mais distante da média real, o viés aumentou, o erro absoluto diminuiu e o RMSE aumentou, devido às previsões do final do gráfico de dispersão.

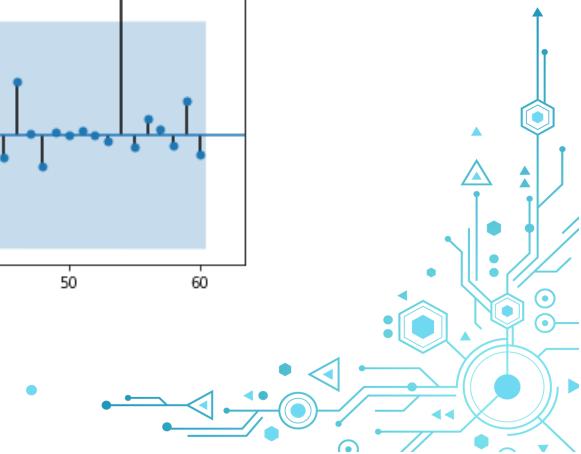
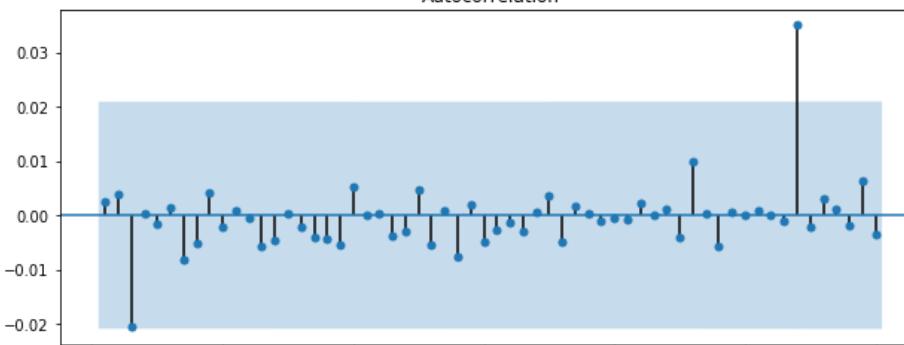
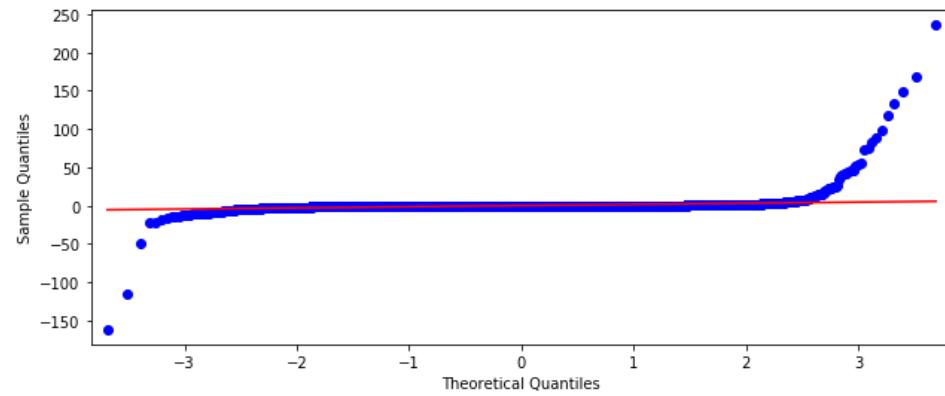
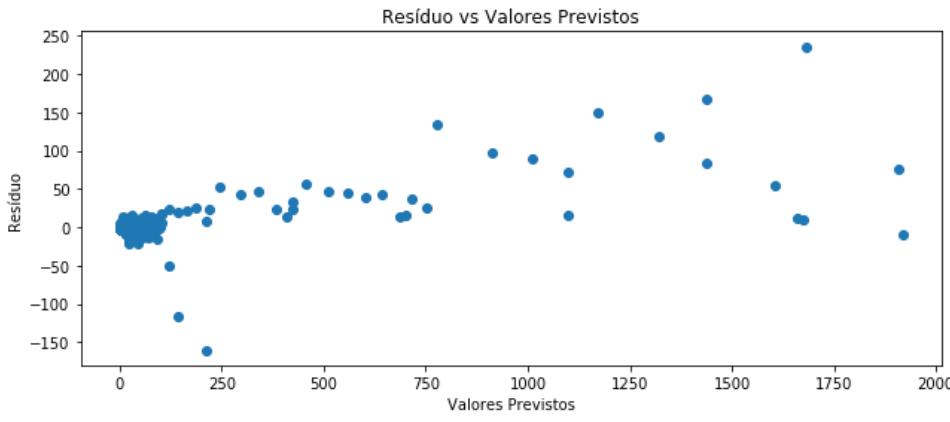
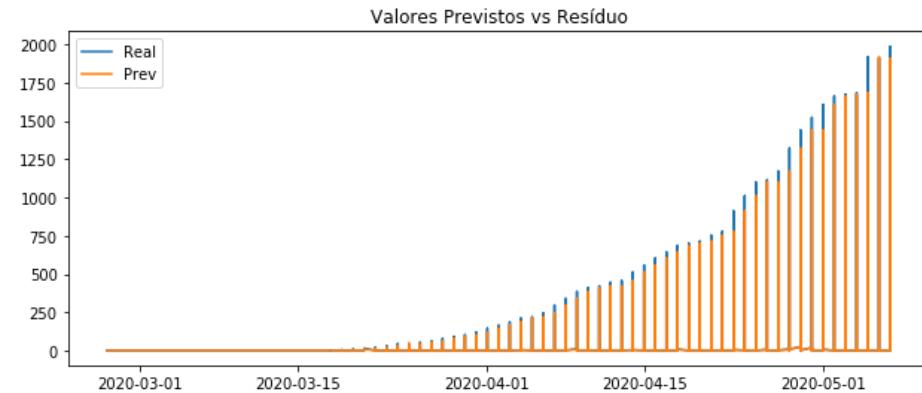


## 5.2.1. Método de Previsão em Laço

ÁRVORE DE DECISÃO | MORTES ACUMULADAS

121

BASE TESTE – COM MODIFICAÇÃO DA DISTRIBUIÇÃO DA TARGET



## 5.2.1. Método de Previsão em Laço

ÁRVORE DE DECISÃO | MORTES ACUMULADAS

122

### Conclusões:

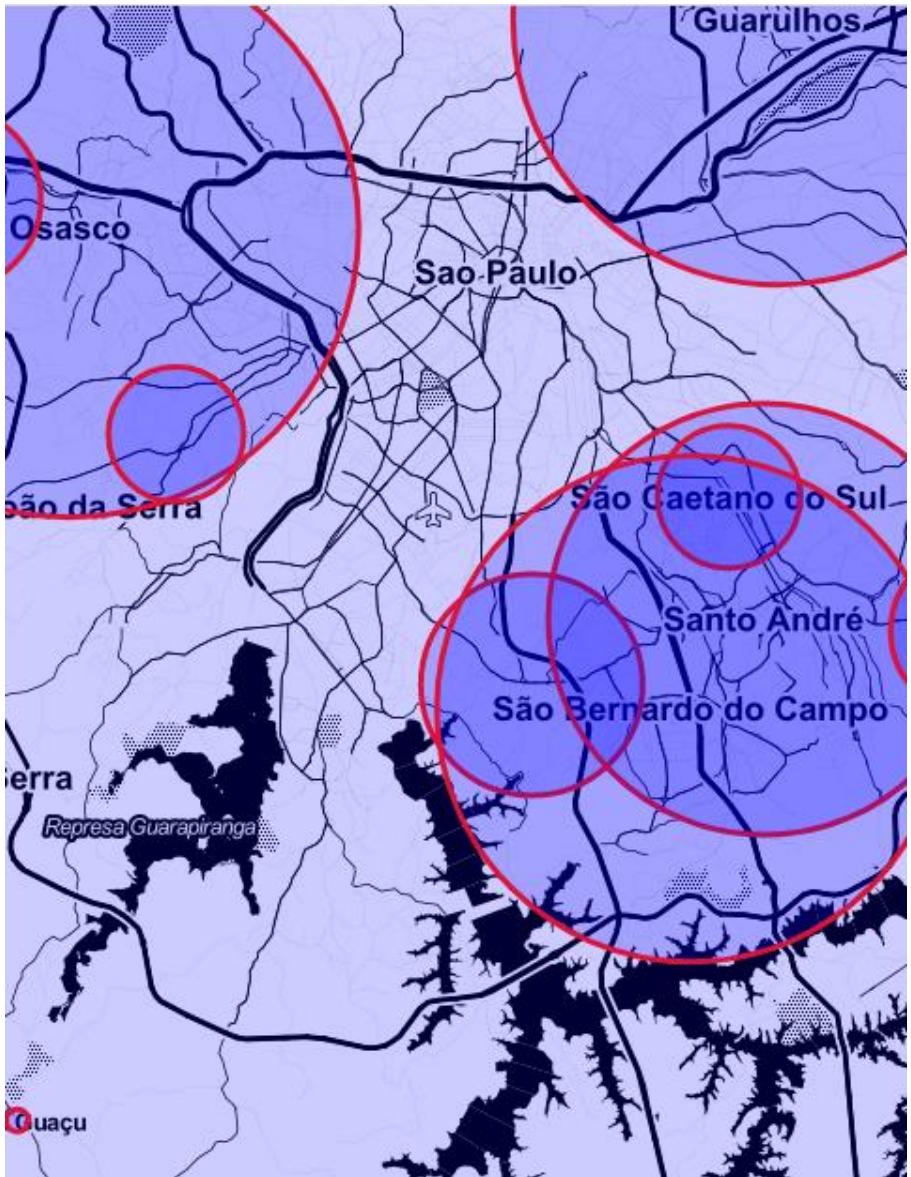
- o método de previsão em laço com árvore de decisão teve performance muito superior à regressão linear para prever mortes acumuladas.
- ainda assim, perdeu por pouco para a baseline.



## 5.2.2. Método de Previsão Tradicional

ÁRVORE DE DECISÃO | CASOS ACUMULADOS

123



## 5.2.2. Método de Previsão Tradicional

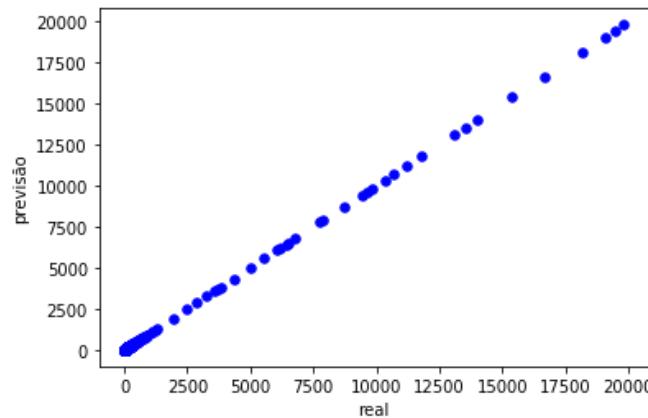
ÁRVORE DE DECISÃO | CASOS ACUMULADOS

124

BASE TREINO – SEM MODIFICAÇÃO DA DISTRIBUIÇÃO DA TARGET

### Árvore de decisão sem modificação da distribuição target:

- Real | Previsto:
  - média: 65.09791 | 65.09791
  - min: 1 | 1
  - 25%: 1 | 1
  - 50%: 3 | 3
  - 75%: 11 | 11
  - max: 19822 | 19822
- sem previsões negativas



Indicadores para base de treino	BASELINE	Árvore de Decisão Sem Modificação da Distr. Target
VIÉS	4.31545	0
MSE	2087.44337	0
RMSE	45.68855	0
MAE	4.41214	0
MAPE	11.90706	0

Com a **base treino**, a árvore de decisão sem modificação da distribuição target **deu overfit total, com 0 erros**.

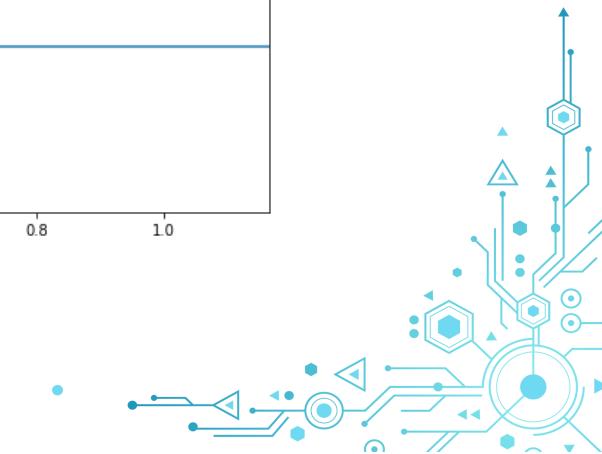
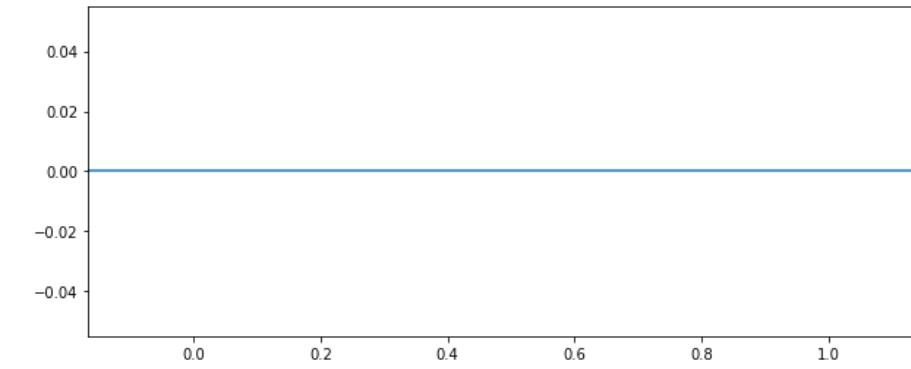
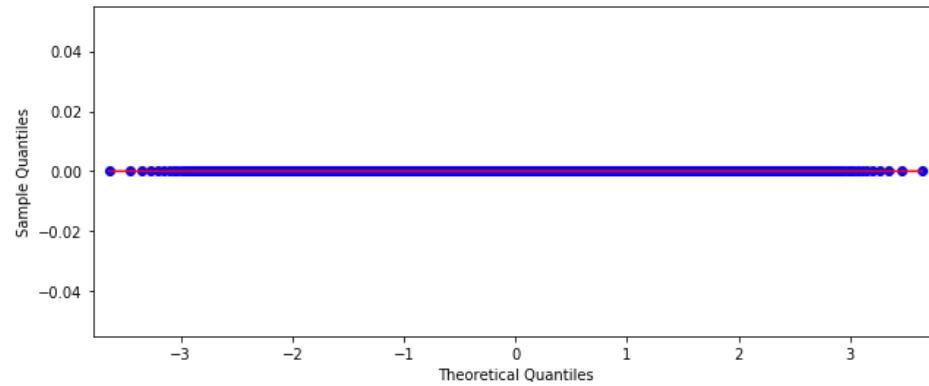
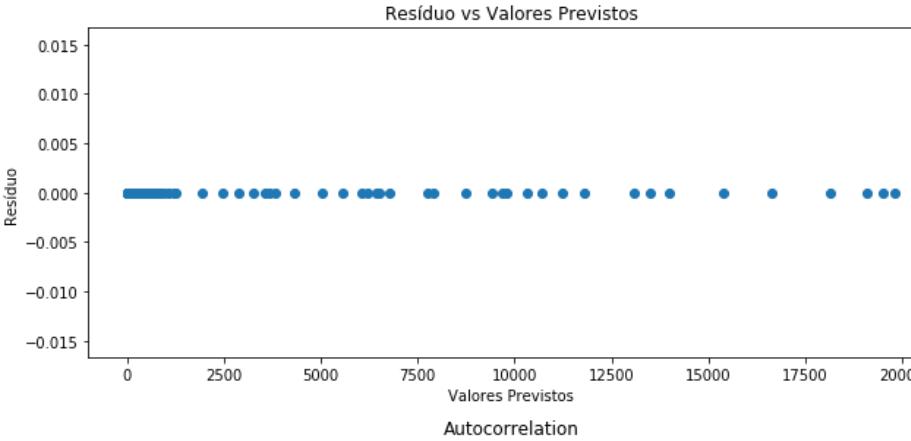
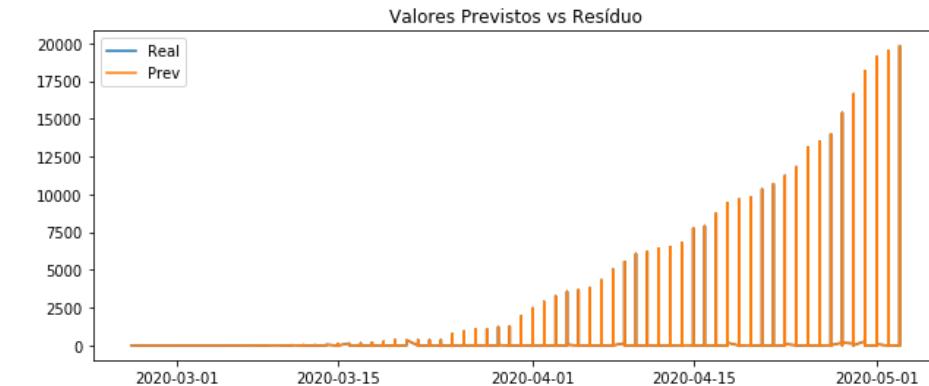


## 5.2.2. Método de Previsão Tradicional

ÁRVORE DE DECISÃO | CASOS ACUMULADOS

125

BASE TREINO – SEM MODIFICAÇÃO DA DISTRIBUIÇÃO DA TARGET



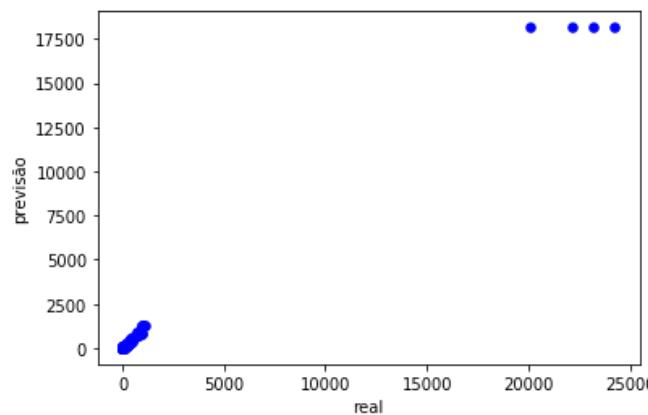
## 5.2.2. Método de Previsão Tradicional

ÁRVORE DE DECISÃO | CASOS ACUMULADOS

126

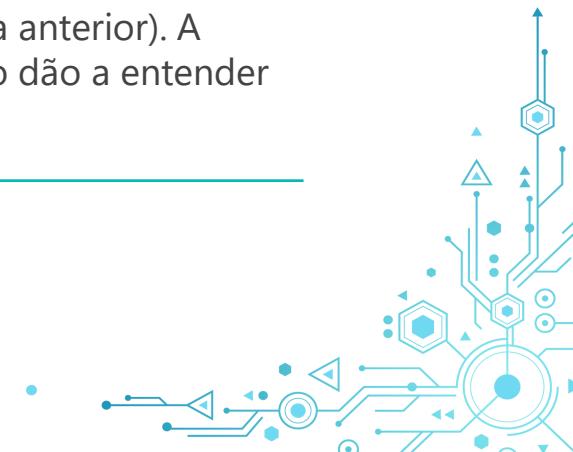
### BASE TESTE – SEM MODIFICAÇÃO DA DISTRIBUIÇÃO DA TARGET

Árvore de decisão sem modificação da distribuição target:
<ul style="list-style-type: none"><li>Real   Previsto:<ul style="list-style-type: none"><li>média: 101.86723   91.12509</li><li>min: 1   1</li><li>25%: 2   2</li><li>50%: 5   5</li><li>75%: 19   19</li><li>max: 24273   18149</li><li>sem previsões negativas</li></ul></li></ul>



Indicadores para base de teste	BASELINE	Árvore de Decisão Sem Modificação da Distr. Target
VIÉS	5.69602	10.74214
MSE	4835.93082	58428.34801
RMSE	69.54086	241.71956
MAE	5.92383	16.42488
MAPE	11.07093	16.48049

Com a **base teste**, a árvore de decisão sem modificação da distribuição target teve **desempenho pior que a baseline**. O erro médio absoluto foi de 16.4 casos, contra 5.9 casos na baseline (que apenas repete o resultado do dia anterior). A principal causa são os outliers da capital, como dão a entender o RMSE e o gráfico de dispersão.

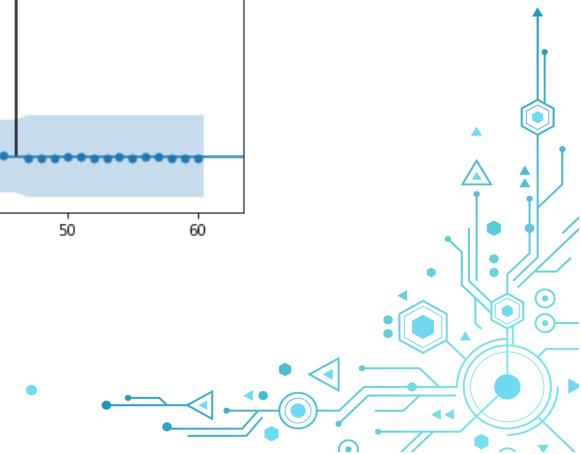
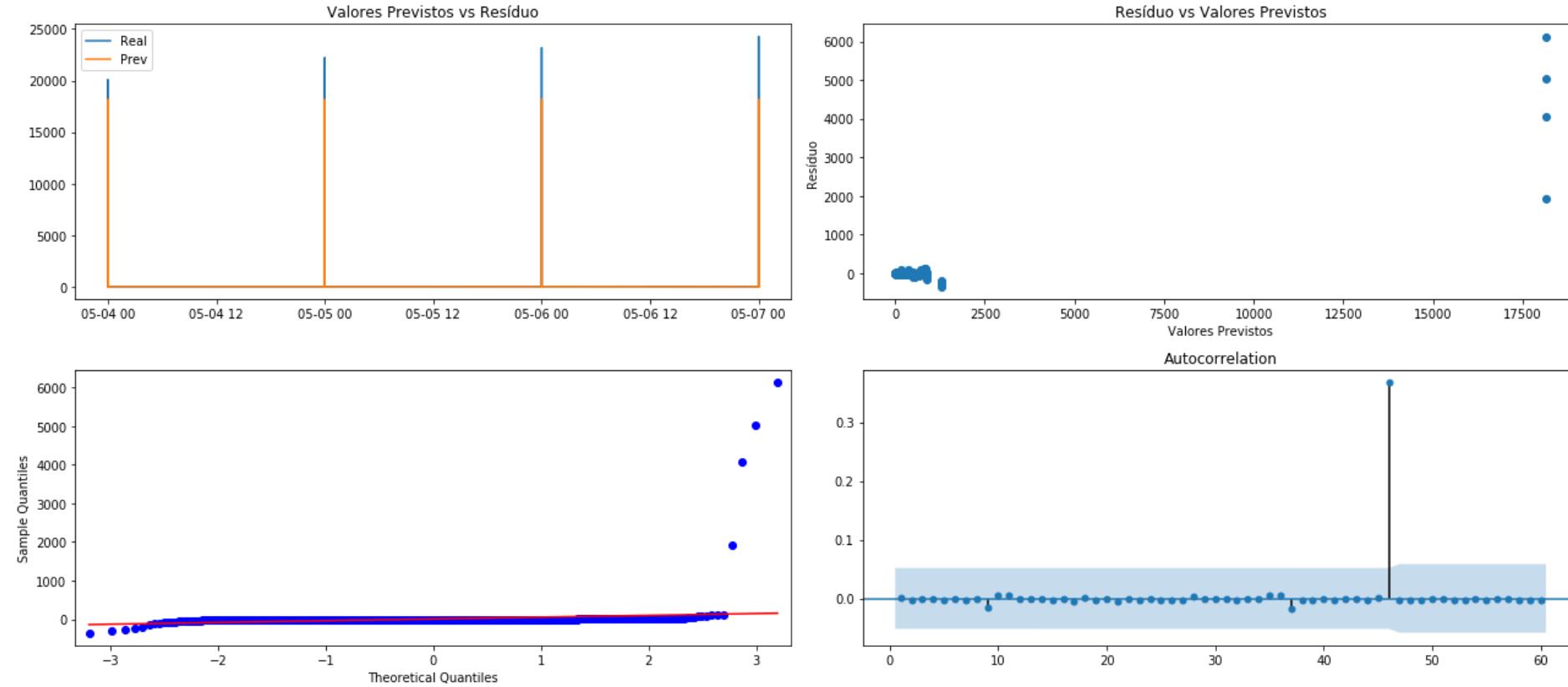


## 5.2.2. Método de Previsão Tradicional

ÁRVORE DE DECISÃO | CASOS ACUMULADOS

127

BASE TESTE – SEM MODIFICAÇÃO DA DISTRIBUIÇÃO DA TARGET



## 5.2.2. Método de Previsão Tradicional

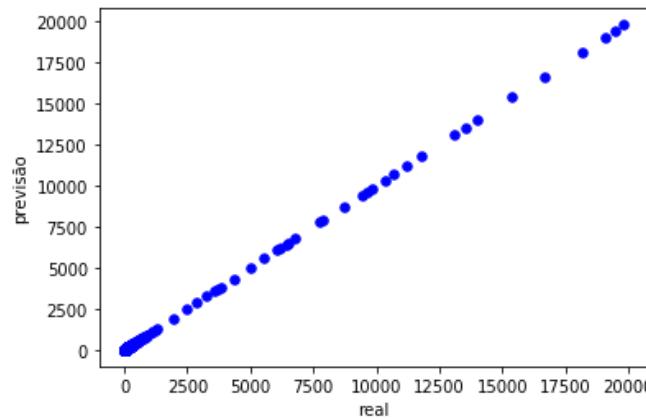
ÁRVORE DE DECISÃO | CASOS ACUMULADOS

128

### BASE TREINO – COM MODIFICAÇÃO DA DISTRIBUIÇÃO DA TARGET

#### Árvore de decisão com modificação da distribuição target:

- Real | Previsto:
  - média: 65.09791 | 65.09791
  - min: 1 | 1
  - 25%: 1 | 1
  - 50%: 3 | 3
  - 75%: 11 | 11
  - max: 19822 | 19822
- sem previsões negativas



Indicadores para base de treino	BASELINE	Árvore de Decisão Com Modificação da Distr. Target
VIÉS	4.31545	0
MSE	2087.44337	0
RMSE	45.68855	0
MAE	4.41214	0
MAPE	11.90706	0

Com a **base treino**, a árvore de decisão com modificação da distribuição target **deu overfit total, com 0 erros**.

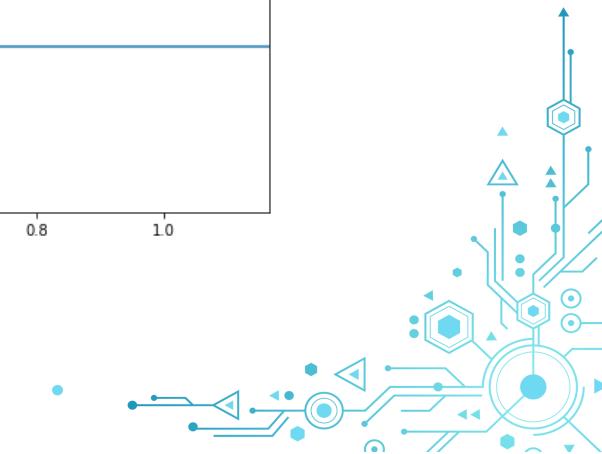
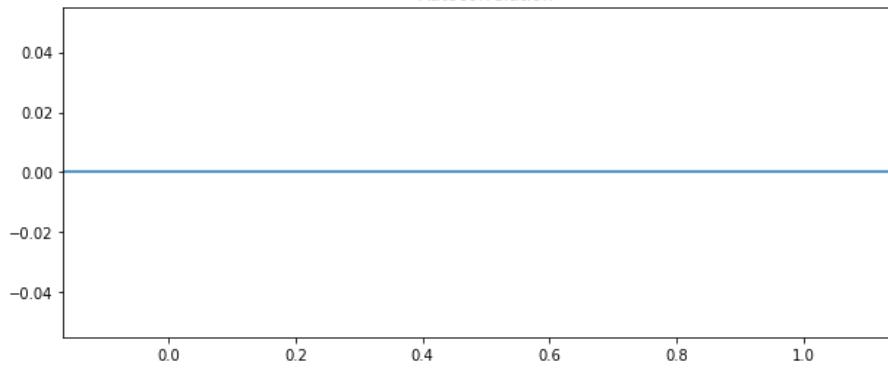
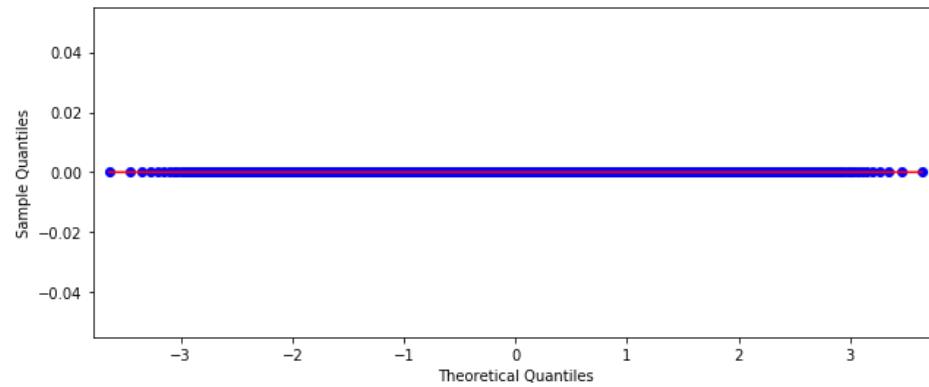
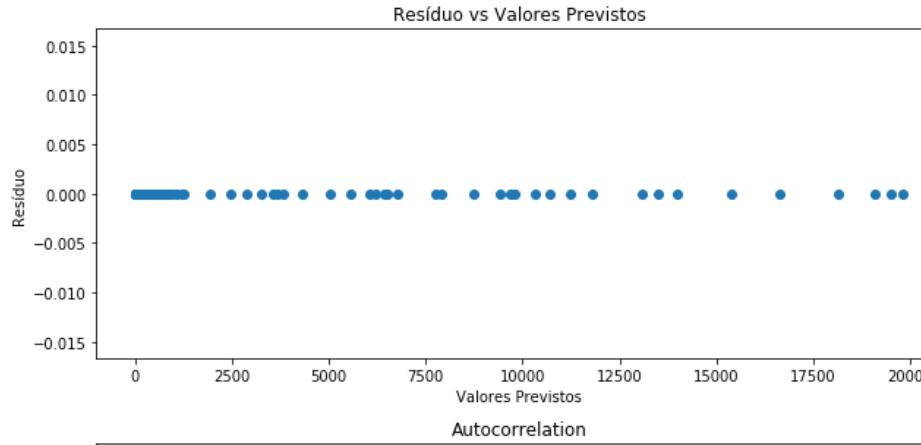
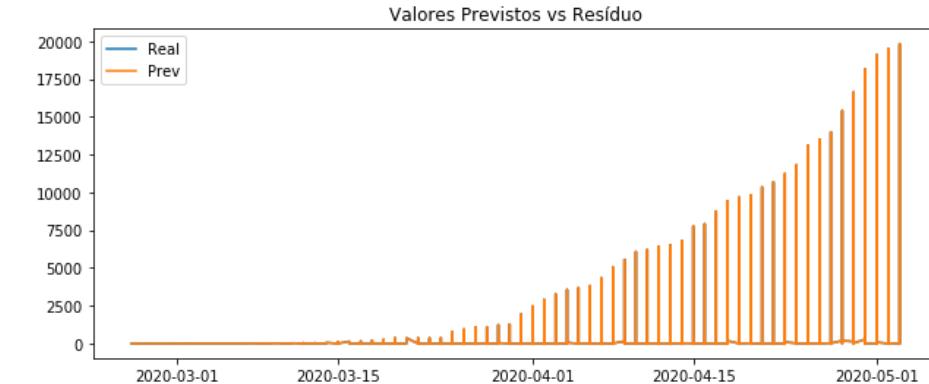


## 5.2.2. Método de Previsão Tradicional

ÁRVORE DE DECISÃO | CASOS ACUMULADOS

129

BASE TREINO – COM MODIFICAÇÃO DA DISTRIBUIÇÃO DA TARGET



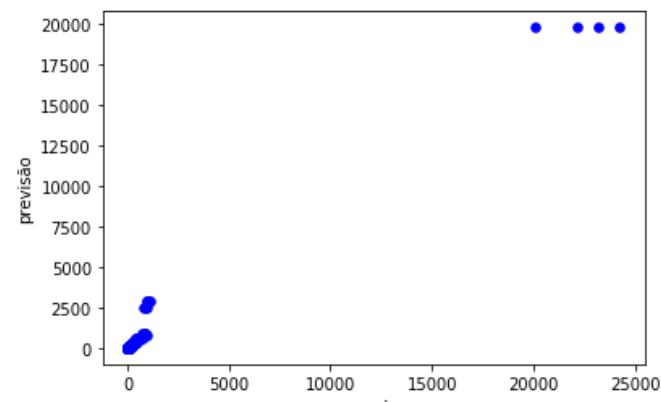
## 5.2.2. Método de Previsão Tradicional

ÁRVORE DE DECISÃO | CASOS ACUMULADOS

130

### BASE TESTE – COM MODIFICAÇÃO DA DISTRIBUIÇÃO DA TARGET

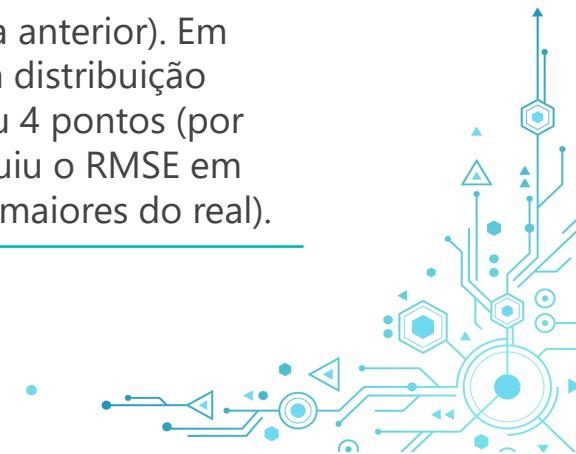
Árvore de decisão com modificação da distribuição target:
<ul style="list-style-type: none"><li>Real   Previsto:<ul style="list-style-type: none"><li>média: 101.86723   104.63871</li><li>min: 1   1</li><li>25%: 2   1</li><li>50%: 5   5</li><li>75%: 19   20</li><li>max: 24273   19822</li><li>sem previsões negativas</li></ul></li></ul>



@2020 LABDATA FIA. Copyright all rights reserved.

Indicadores para base de teste	BASELINE	Árvore de Decisão Com Modificação da Distr. Target
VIÉS	5.69602	-2.77149
MSE	4835.93082	43586.11461
RMSE	69.54086	208.77288
MAE	5.92383	20.59259
MAPE	11.07093	16.46912

Com a **base teste**, a árvore de decisão com modificação da distribuição target teve **desempenho pior que a baseline**. O erro médio absoluto foi de 20.5 casos, contra 5.9 casos na baseline (que apenas repete o resultado do dia anterior). Em comparação com a árvore sem modificação da distribuição target, vemos que o erro médio absoluto subiu 4 pontos (por causa dos erros no 2º e 3º quartis), mas diminuiu o RMSE em mais de 30 pontos (aproximando as previsões maiores do real).

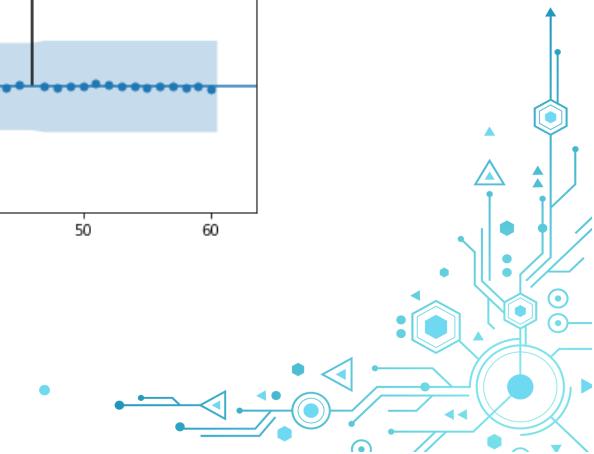
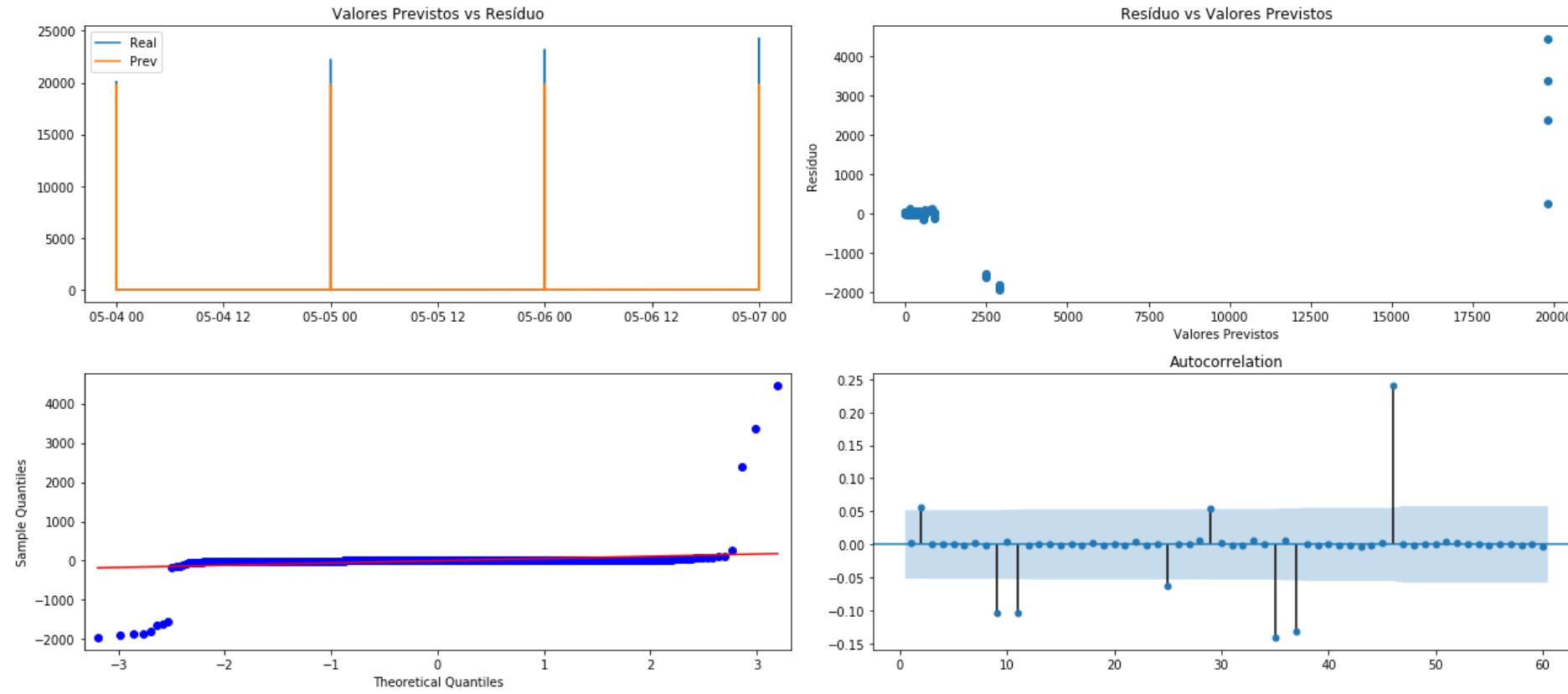


## 5.2.2. Método de Previsão Tradicional

ÁRVORE DE DECISÃO | CASOS ACUMULADOS

131

BASE TESTE – COM MODIFICAÇÃO DA DISTRIBUIÇÃO DA TARGET



## 5.2.2. Método de Previsão Tradicional

ÁRVORE DE DECISÃO | CASOS ACUMULADOS

132

### Conclusões:

- o método de previsão tradicional com árvore de decisão teve performance muito superior à regressão linear para prever casos acumulados.
- ainda assim, perdeu para a baseline e para o método de previsão em laço.



## 5.2.2. Método de Previsão Tradicional

ÁRVORE DE DECISÃO | MORTES ACUMULADAS

133



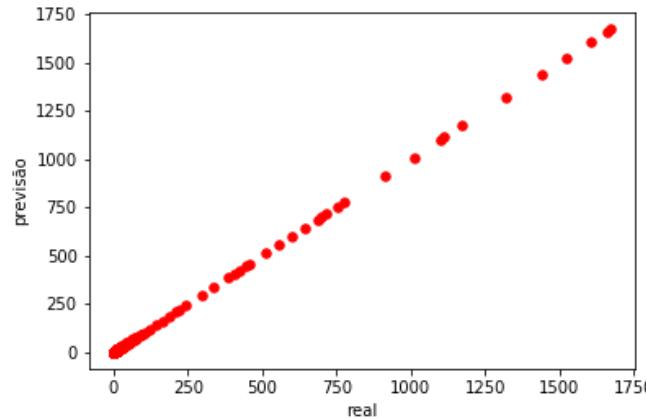
## 5.2.2. Método de Previsão Tradicional

ÁRVORE DE DECISÃO | MORTES ACUMULADAS

134

BASE TREINO – SEM MODIFICAÇÃO DA DISTRIBUIÇÃO DA TARGET

Árvore de decisão sem modificação da distribuição target:
• Real   Previsto:
• média: 4.98669   4.98669
• min: 0   0
• 25%: 0   0
• 50%: 0   0
• 75%: 1   1
• max: 1673   1673
• sem previsões negativas



Indicadores para base de treino	BASELINE	Árvore de Decisão Sem Modificação da Distr. Target
VIÉS	0.35646	0
MSE	16.43061	0
RMSE	4.05347	0
MAE	0.36353	0
MAPE	9.64655	0

Com a **base treino**, a árvore de decisão sem modificação da distribuição target **deu overfit total, com 0 erros**.

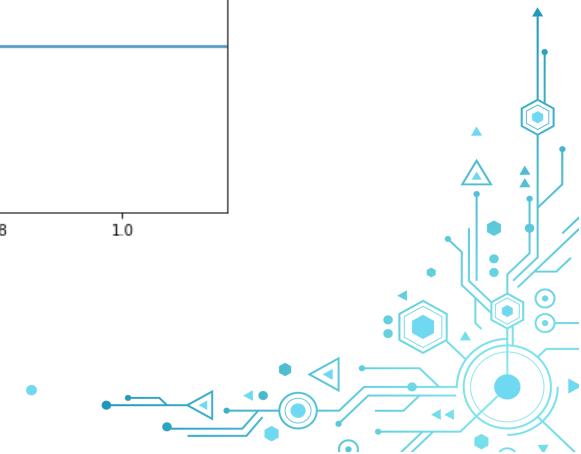
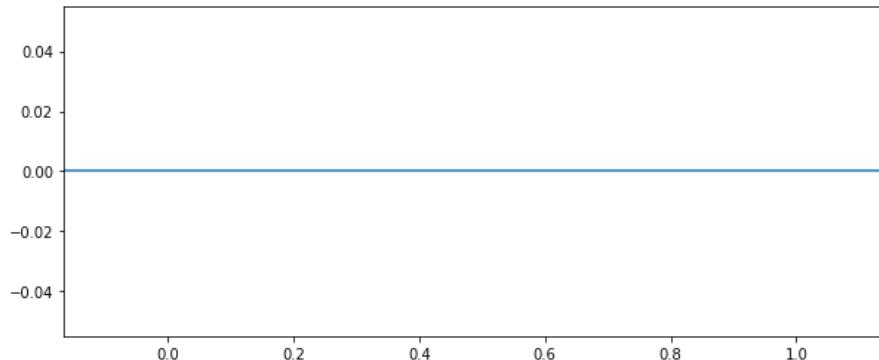
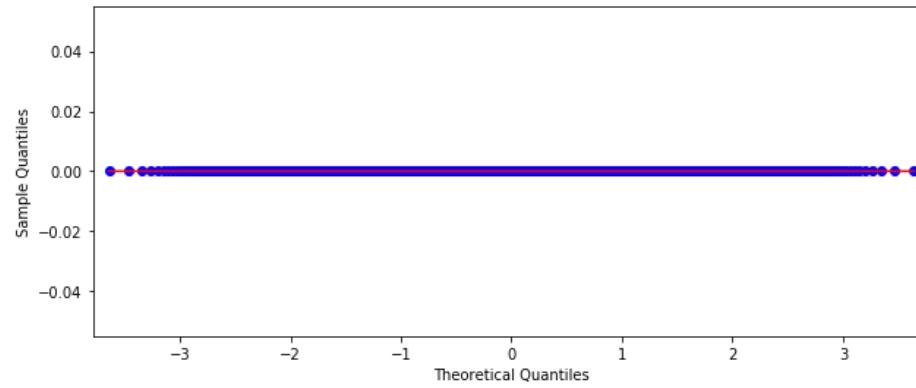
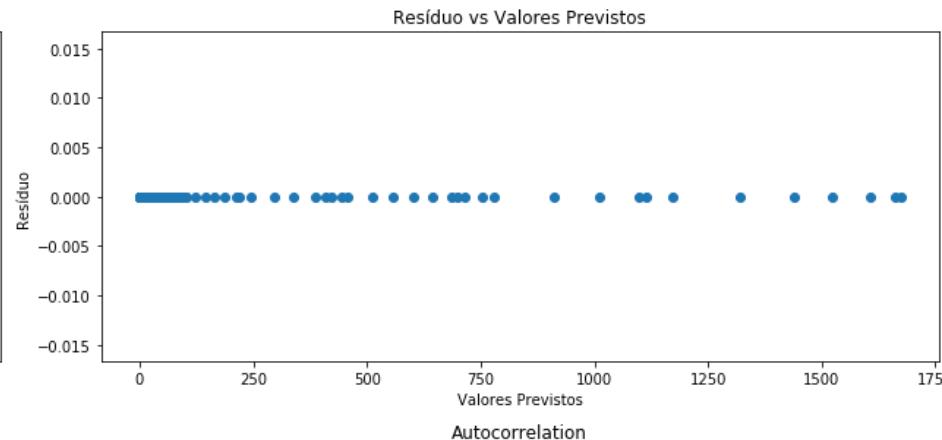
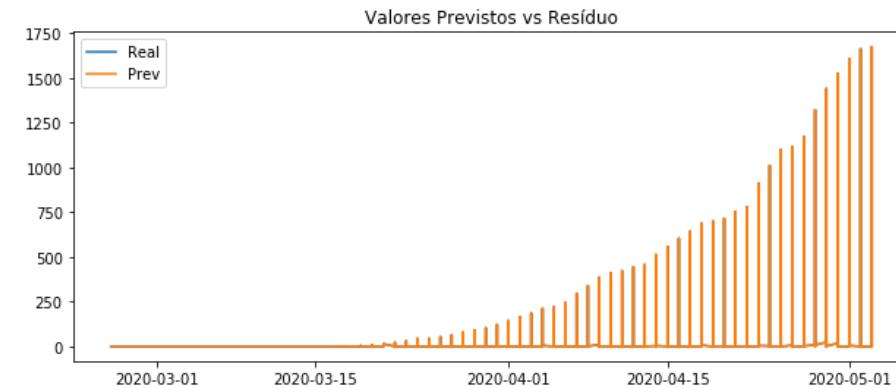


## 5.2.2. Método de Previsão Tradicional

ÁRVORE DE DECISÃO | MORTES ACUMULADAS

135

BASE TREINO – SEM MODIFICAÇÃO DA DISTRIBUIÇÃO DA TARGET



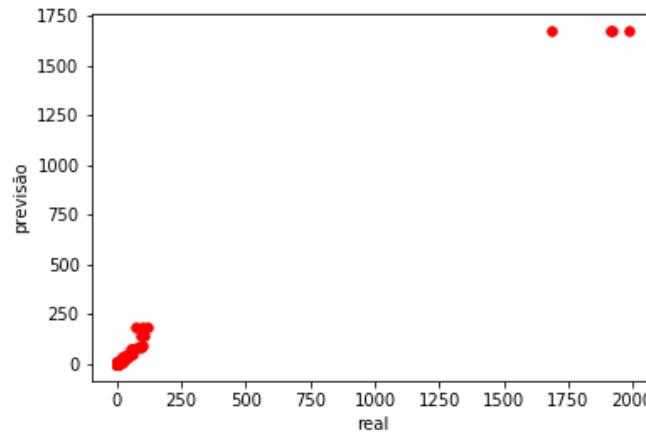
## 5.2.2. Método de Previsão Tradicional

ÁRVORE DE DECISÃO | MORTES ACUMULADAS

136

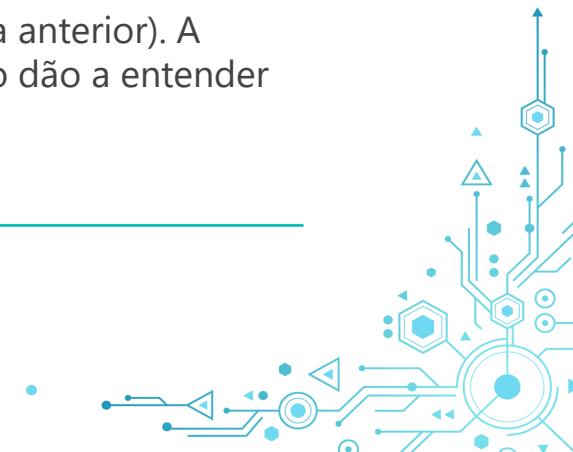
### BASE TESTE – SEM MODIFICAÇÃO DA DISTRIBUIÇÃO DA TARGET

Árvore de decisão sem modificação da distribuição target:
• Real   Previsto:
• média: 8.34731   8.22292
• min: 0   0
• 25%: 0   0
• 50%: 0   0
• 75%: 2   2
• max: 1986   1673
• sem previsões negativas



Indicadores para base de teste	BASELINE	Árvore de Decisão Sem Modificação da Distr. Target
VIÉS	0.40601	0.12439
MSE	44.75681	172.75611
RMSE	6.69005	13.14367
MAE	0.47729	1.24808
MAPE	6.72089	inf

Com a **base teste**, a árvore de decisão sem modificação da distribuição target teve **desempenho pior que a baseline**. O erro médio absoluto foi de 1.24 mortes, contra 0.47 mortes na baseline (que apenas repete o resultado do dia anterior). A principal causa são os outliers da capital, como dão a entender o RMSE e o gráfico de dispersão.

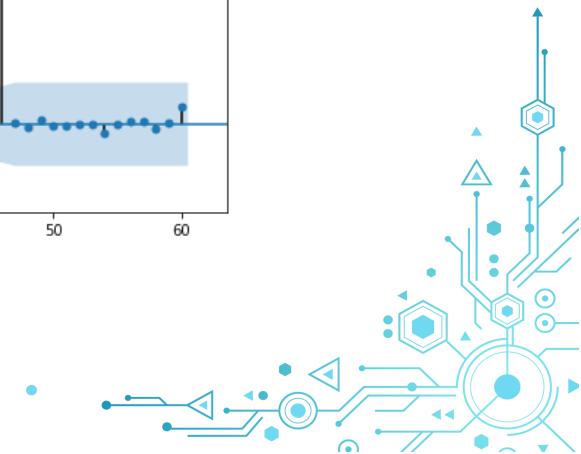
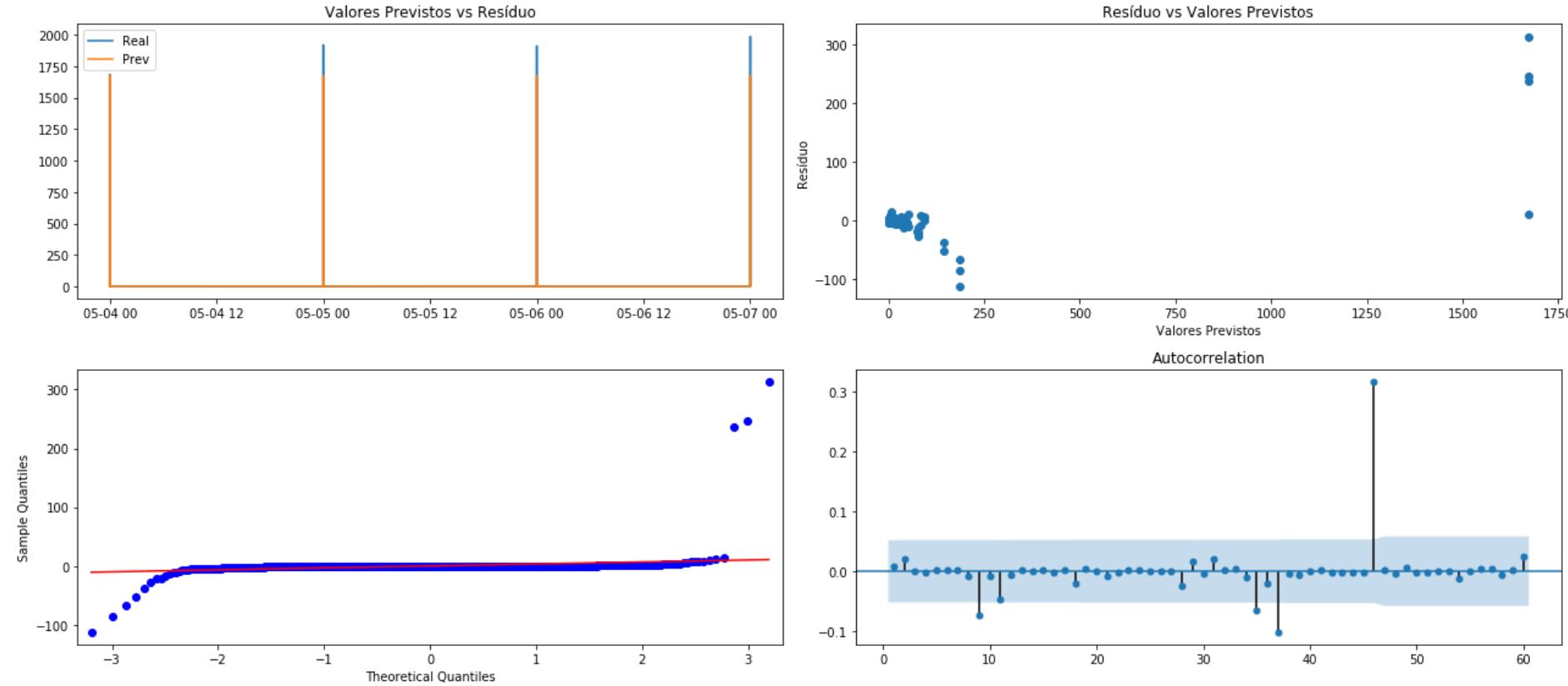


## 5.2.2. Método de Previsão Tradicional

ÁRVORE DE DECISÃO | MORTES ACUMULADAS

137

BASE TESTE – SEM MODIFICAÇÃO DA DISTRIBUIÇÃO DA TARGET



## 5.2.2. Método de Previsão Tradicional

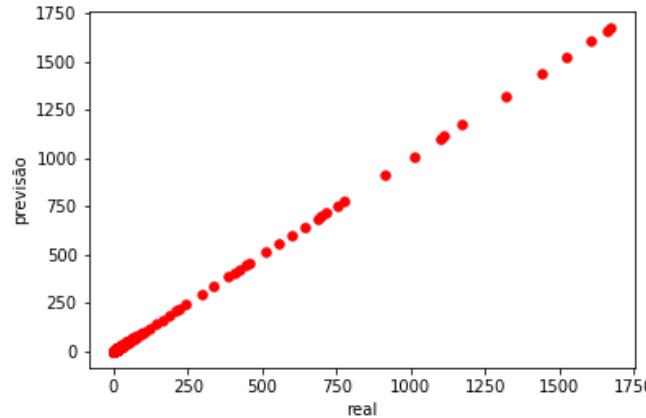
ÁRVORE DE DECISÃO | MORTES ACUMULADAS

138

### BASE TREINO – COM MODIFICAÇÃO DA DISTRIBUIÇÃO DA TARGET

#### Árvore de decisão com modificação da distribuição target:

- Real | Previsto:
  - média: 4.98669 | 4.98669
  - min: 0 | 0
  - 25%: 0 | 0
  - 50%: 0 | 0
  - 75%: 1 | 1
  - max: 1673 | 1673
- sem previsões negativas



Indicadores para base de treino	BASELINE	Árvore de Decisão Com Modificação da Distr. Target
VIÉS	0.35646	0
MSE	16.43061	0
RMSE	4.05347	0
MAE	0.36353	0
MAPE	9.64655	0

Com a **base treino**, a árvore de decisão com modificação da distribuição target **deu overfit total, com 0 erros**.

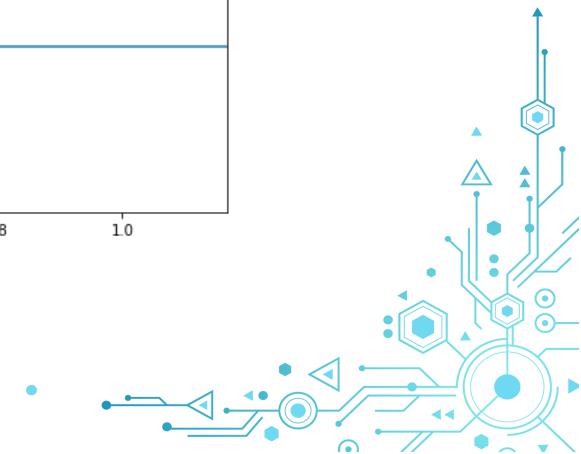
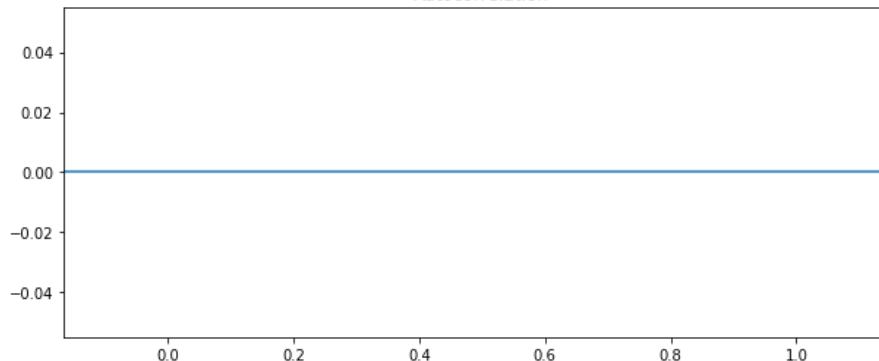
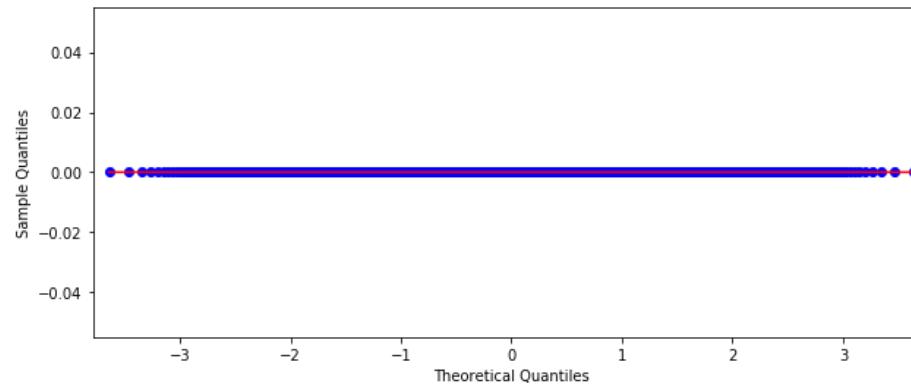
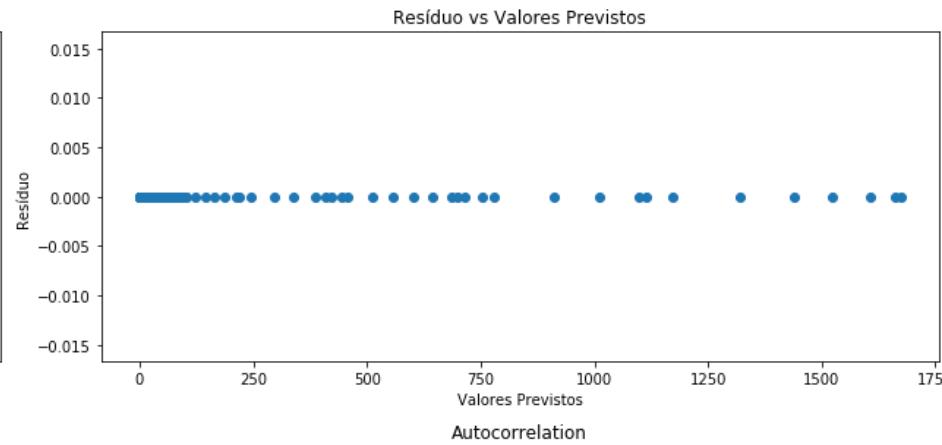
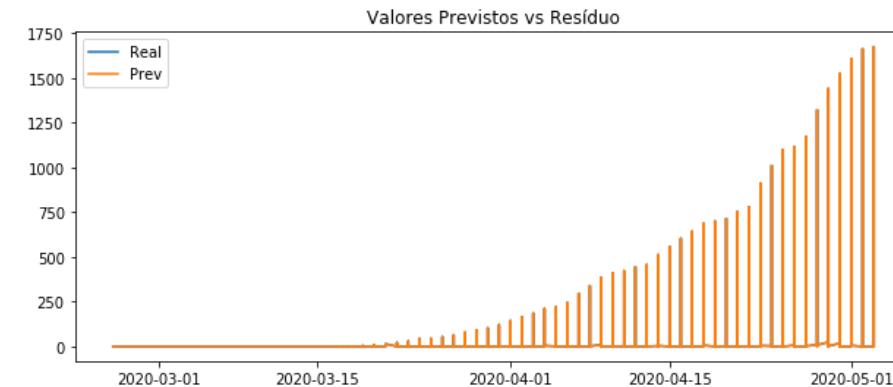


## 5.2.2. Método de Previsão Tradicional

ÁRVORE DE DECISÃO | MORTES ACUMULADAS

139

BASE TREINO – COM MODIFICAÇÃO DA DISTRIBUIÇÃO DA TARGET



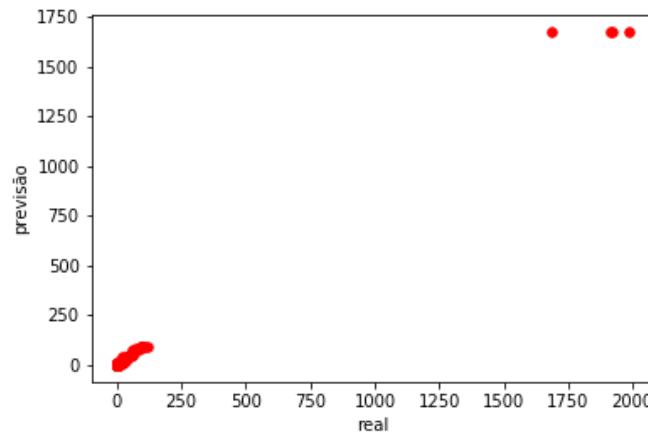
## 5.2.2. Método de Previsão Tradicional

ÁRVORE DE DECISÃO | MORTES ACUMULADAS

140

### BASE TESTE – COM MODIFICAÇÃO DA DISTRIBUIÇÃO DA TARGET

Árvore de decisão com modificação da distribuição target:
• Real   Previsto:
• média: 8.34731   7.87771
• min: 0   0
• 25%: 0   0
• 50%: 0   0
• 75%: 2   2
• max: 1986   1673
• sem previsões negativas



Indicadores para base de teste	BASELINE	Árvore de Decisão Com Modificação da Distr. Target
VIÉS	0.40601	0.46960
MSE	44.75681	153.15164
RMSE	6.69005	12.37545
MAE	0.47729	1.03424
MAPE	6.72089	inf

Com a **base teste**, a árvore de decisão com modificação da distribuição target teve **desempenho pior que a baseline**. O erro médio absoluto foi de 1.03 mortes, contra 0.47 mortes na baseline (que apenas repete o resultado do dia anterior). Em comparação com a árvore sem modificação da distribuição target, vemos que o erro médio absoluto caiu (canto inferior do gráfico de dispersão), assim como o RMSE.

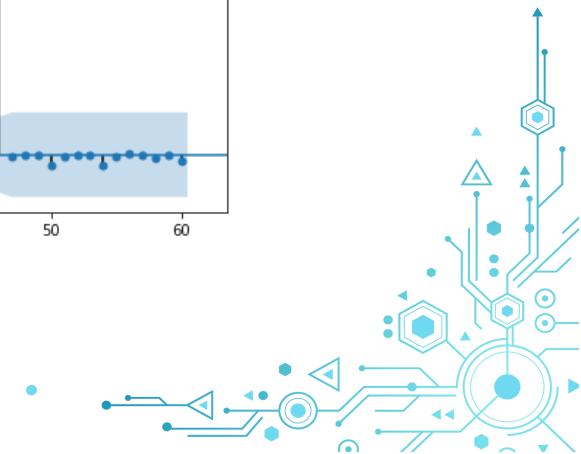
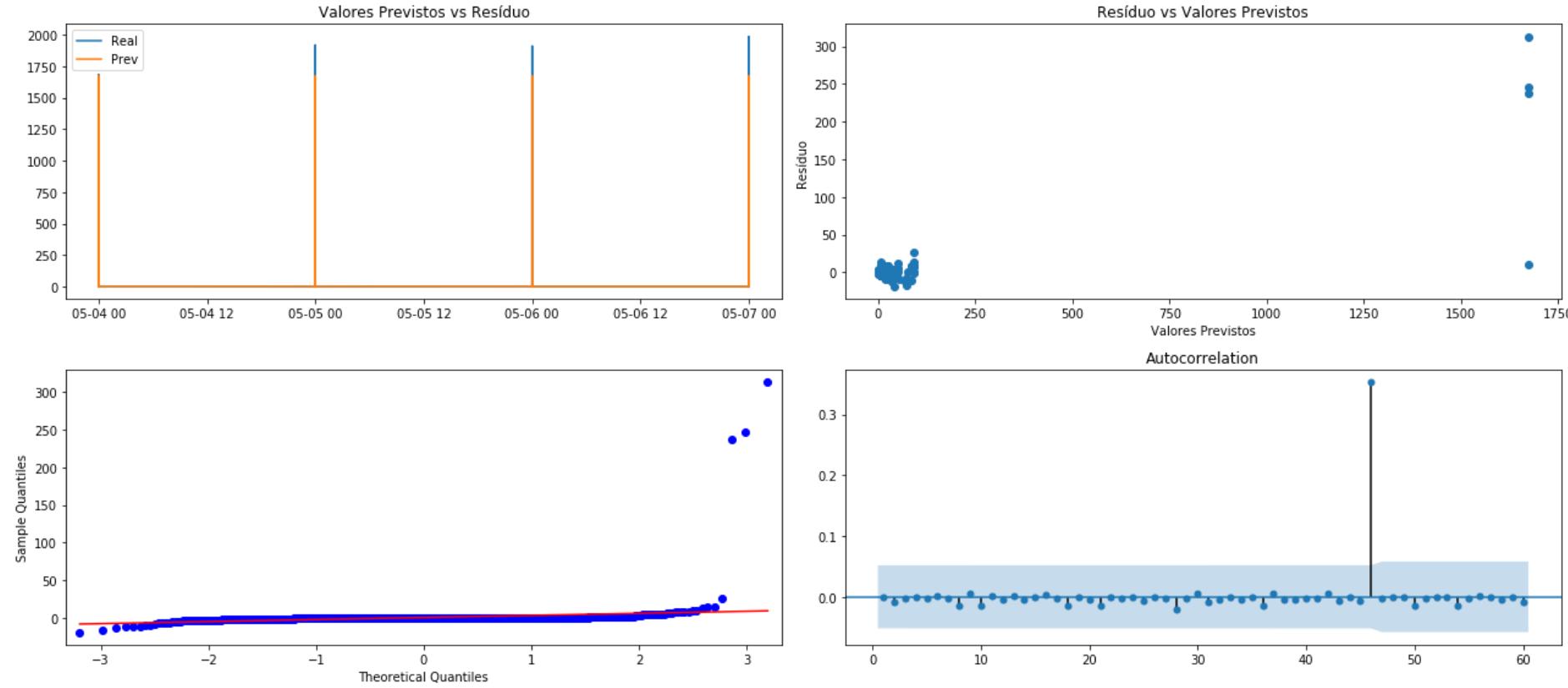


## 5.2.2. Método de Previsão Tradicional

ÁRVORE DE DECISÃO | MORTES ACUMULADAS

141

BASE TESTE – COM MODIFICAÇÃO DA DISTRIBUIÇÃO DA TARGET



## 5.2.2. Método de Previsão Tradicional

ÁRVORE DE DECISÃO | MORTES ACUMULADAS

142

### Conclusões:

- o método de previsão tradicional com árvore de decisão teve performance muito superior à regressão linear para prever mortes acumuladas.
- ainda assim, perdeu para a baseline e para o método de previsão em laço.



# Metodologia de Análise de Dados

143



# 6. Modelagem com Inteligência Artificial

MODELOS UTILIZADOS | ALGORITMOS DE INTELIGÊNCIA ARTIFICIAL

144

Para **superar as métricas da baseline**, partimos para a abordagem do **aprendizado de máquina**, já que as técnicas de estatística tradicional não tiveram o desempenho desejado.

Trabalhamos inicialmente com 3 algoritmos diferentes e aplicamos em todos a técnica de grid search para encontrar os melhores hiperparâmetros.

## Algoritmos de Machine Learning

Árvore de Decisão

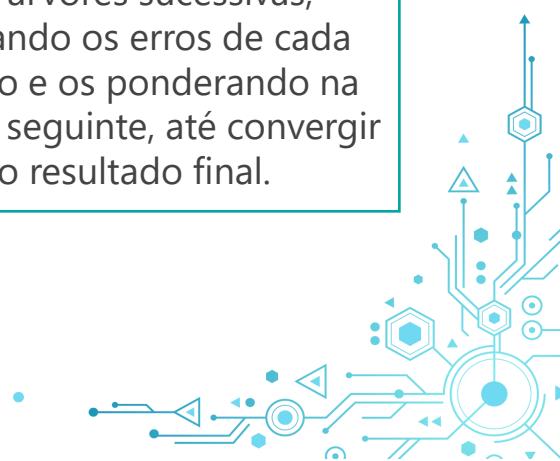
Problemas complexos são quebrados em subproblemas menos complexos, de forma recursiva, formando uma árvore.

Random Forest

Cria árvores aleatórias para amostras aleatórias do dataset. Em seguida, realiza uma votação para estabelecer a regra vencedora.

Gradient Boosting

Cria árvores sucessivas, calculando os erros de cada iteração e os ponderando na iteração seguinte, até convergir no resultado final.



# 6. Modelagem com Inteligência Artificial

FEATURE ENGINEERING | SELEÇÃO DE VARIÁVEIS

145

Aplicamos 3 métodos de seleção de variáveis:

- Baseado em filtro ( $\chi^2$ )
- Baseado em wrapper (regressão linear)
- Baseado em método embarcado (random forest regressor)



Na comparação dos resultados de cada método, optamos por:

- **53** variáveis selecionadas em pelo menos 1 método para previsão de **casos acumulados** com árvore de decisão.
- **47** variáveis selecionadas em pelo menos 1 método para previsão de **mortes acumuladas** com árvore de decisão.



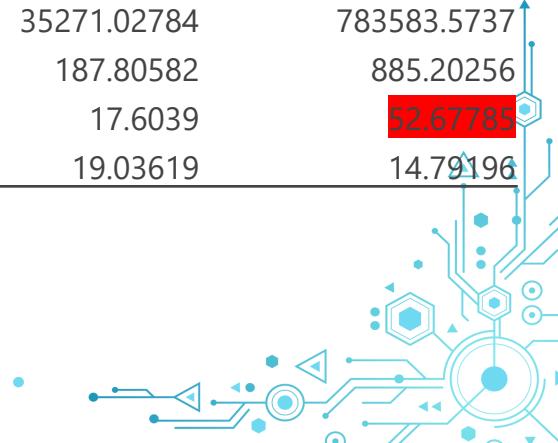
# 6. Modelagem com Inteligência Artificial

COMPARAÇÃO DE MODELOS | RESULTADOS: CASOS ACUMULADOS

146

Base Teste - Casos Acumulados Método de Previsão em Laço	Baseline	Árvore de Decisão Sem Modificação da Distr. Target	Árvore de Decisão Com Modificação da Distr. Target	Random Forest Sem Modificação da Distr. Target	Random Forest Com Modificação da Distr. Target	Gradient Boosting Sem Modificação da Distr. Target	Gradient Boosting Com Modificação da Distr. Target
VIÉS	4.54008	1.91506	15.1816	1.66887	3.41653	0.26325	27.64817
MSE	2534.639	3027.51228	93377.15431	2816.98453	6331.00398	24725.20184	287254.1524
RMSE	50.3452	55.02283	305.57676	53.07527	79.56761	157.24249	535.96096
MAE	4.6581	5.82272	23.01717	5.67432	7.42199	17.04105	31.64999
MAPE	11.77102	13.69047	12.81933	13.45639	11.18903	24.39874	15.56567

Base Teste - Casos Acumulados Método de Previsão Tradicional	Baseline	Árvore de Decisão Sem Modificação da Distr. Target	Árvore de Decisão Com Modificação da Distr. Target	Random Forest Sem Modificação da Distr. Target	Random Forest Com Modificação da Distr. Target	Gradient Boosting Sem Modificação da Distr. Target	Gradient Boosting Com Modificação da Distr. Target
VIÉS	5.69602	-1.0877	3.82739	3.24273	3.59119	2.34562	46.10203
MSE	4835.93082	45650.44532	160320.202	34803.3944	36779.68484	35271.02784	783583.5737
RMSE	69.54086	213.65965	400.40005	186.55668	191.7803	187.80582	885.20256
MAE	5.92383	18.80957	35.25017	15.27726	15.44165	17.6039	52.67785
MAPE	11.07093	13.17174	10.66033	14.40768	10.30301	19.03619	14.79196



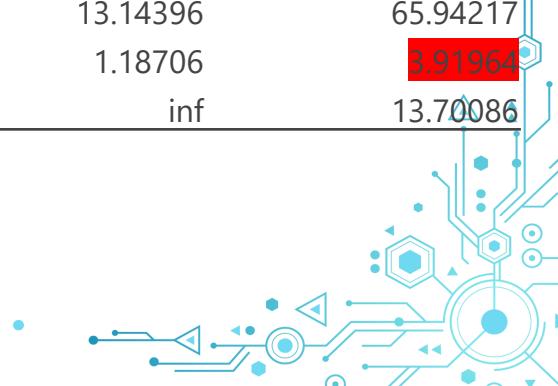
# 6. Modelagem com Inteligência Artificial

COMPARAÇÃO DE MODELOS | RESULTADOS: MORTES ACUMULADAS

147

Base Teste – Mortes Acumuladas Método de Previsão em Laço	Baseline	Árvore de Decisão Sem Modificação da Distr. Target	Árvore de Decisão Com Modificação da Distr. Target	Random Forest Sem Modificação da Distr. Target	Random Forest Com Modificação da Distr. Target	Gradient Boosting Sem Modificação da Distr. Target	Gradient Boosting Com Modificação da Distr. Target
VIÉS	0.36453	0.14794	1.46327	0.16898	0.3731	0.18933	2.23038
MSE	21.03945	26.58347	616.1053	21.90721	45.94485	24.07767	1699.40573
RMSE	4.58688	5.15592	24.82147	4.68051	6.77826	4.9069	41.22385
MAE	0.38204	0.56709	1.76461	0.48556	0.60075	0.5589	2.48465
MAPE	nan	inf	inf	inf	inf	inf	inf

Base Teste - Mortes Acumuladas Método de Previsão Tradicional	Baseline	Árvore de Decisão Sem Modificação da Distr. Target	Árvore de Decisão Com Modificação da Distr. Target	Random Forest Sem Modificação da Distr. Target	Random Forest Com Modificação da Distr. Target	Gradient Boosting Sem Modificação da Distr. Target	Gradient Boosting Com Modificação da Distr. Target
VIÉS	0.40601	0.30219	1.52131	0.1787	0.37456	0.23741	3.45283
MSE	44.75681	164.681	728.16981	163.43438	156.90985	172.7636	4348.36967
RMSE	6.69005	12.83281	26.98462	12.78415	12.52637	13.14396	65.94217
MAE	0.47729	1.13552	1.79525	1.15826	0.98532	1.18706	3.91964
MAPE	6.72089	inf	8.85452	inf	9.32019	inf	13.70086



# 6. Modelagem com Inteligência Artificial

COMPARAÇÃO DE MODELOS | RESULTADOS

148

## Conclusões:

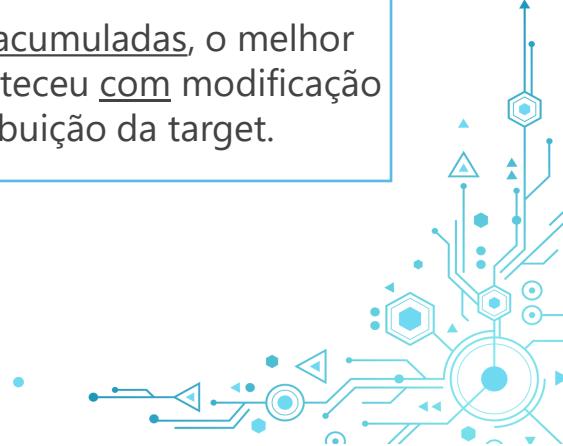
Nenhum modelo conseguiu superar a baseline.

O método de previsão em laço teve performance melhor do que o método tradicional, pois ele é treinado com uma base maior a cada iteração.

A técnica de Random Forest foi a que chegou mais perto da baseline.

Para casos acumulados, o melhor resultado aconteceu sem modificar a distribuição da target.

Para mortes acumuladas, o melhor resultado aconteceu com modificação da distribuição da target.



# 6.1. Modelagem com Inteligência Artificial

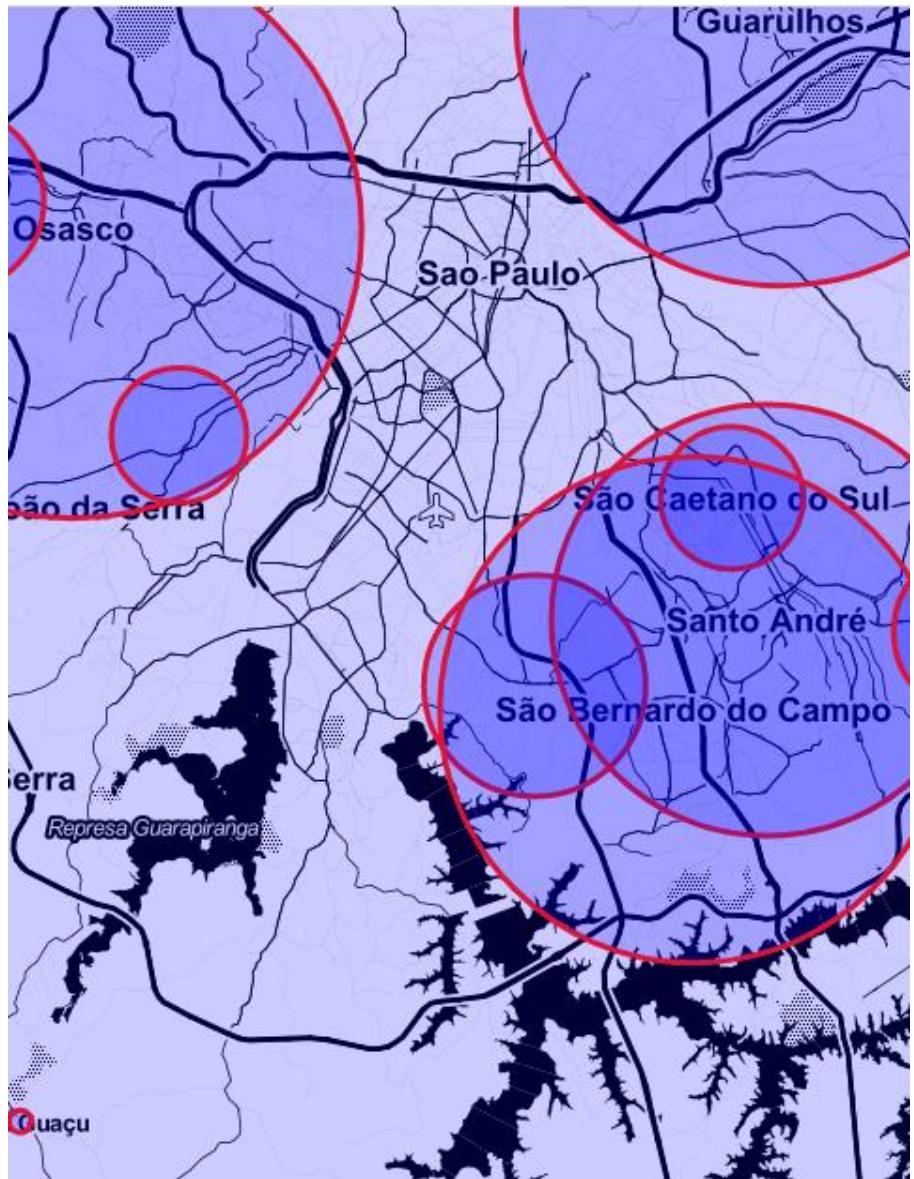
APLICAÇÃO DE MODELO | RANDOM FOREST

149

## RANDOM FOREST

## 6.1. Método de Previsão em Laço

RANDOM FOREST | CASOS ACUMULADOS



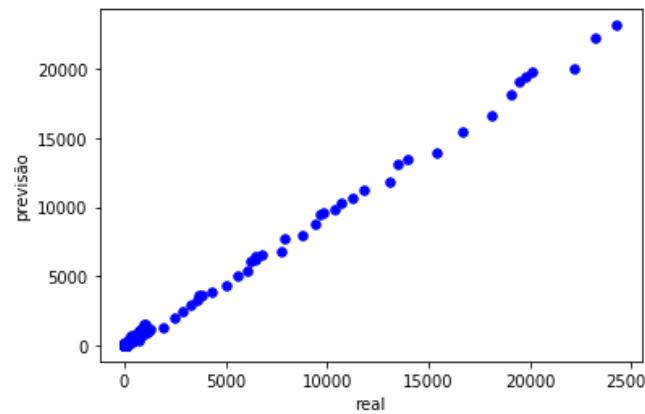
## 6.1. Método de Previsão em Laço

## RANDOM FOREST | CASOS ACUMULADOS

## **BASE TESTE – SEM MODIFICAÇÃO DA DISTRIBUIÇÃO DA TARGET**

## **Random Forest sem modificação da distribuição target:**

- Real | Previsto:
    - Média: 71.08847 | 69.41960
    - min: 1 | 1
    - 25%: 1 | 1
    - 50%: 3 | 3
    - 75%: 13 | 13
    - max: 24273 | 23158
    - sem previsões negativas



<b>Indicadores para base de teste</b>	<b>BASELINE</b>	<b>Random Forest (laço)</b>
VIÉS	4.54008	1.66887
MSE	2534.639	2816.98453
RMSE	50.3452	53.07527
MAE	4.6581	5.67432
MAPE	11.77102	13.45639

Com random forest sem modificação da distribuição da target, o laço funcionou durante todos os **128 dias**, com indicadores de erros aumentando gradativamente.

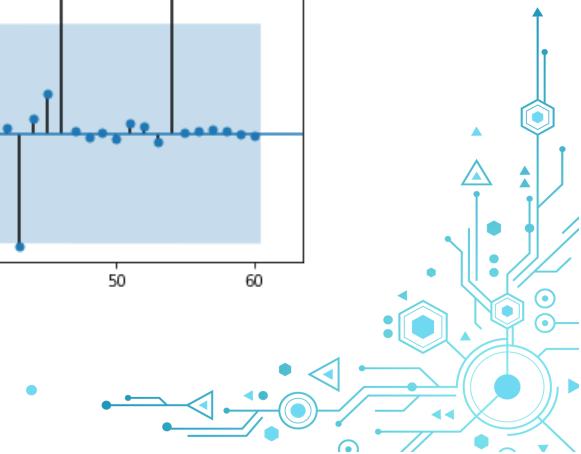
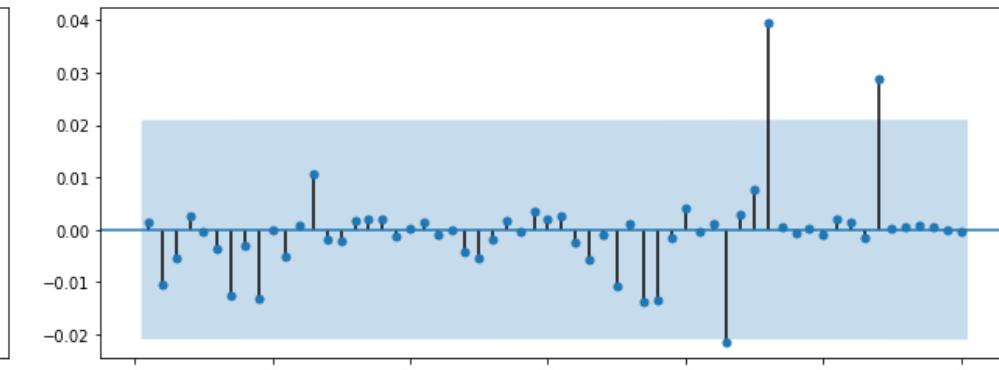
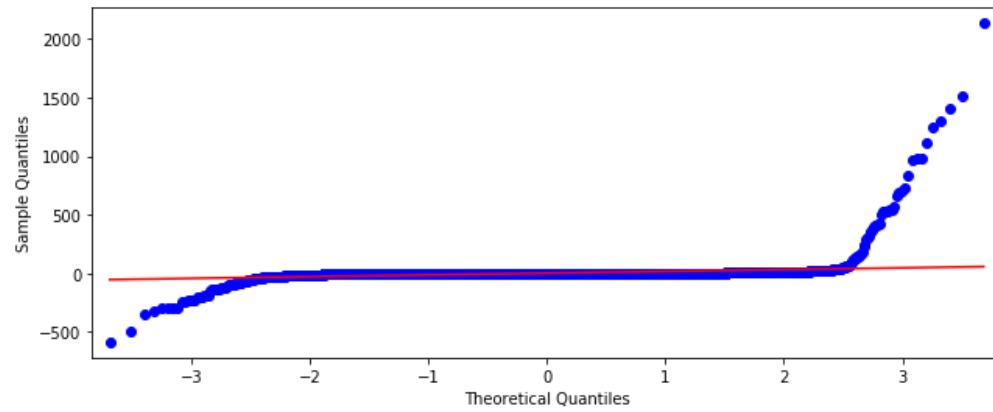
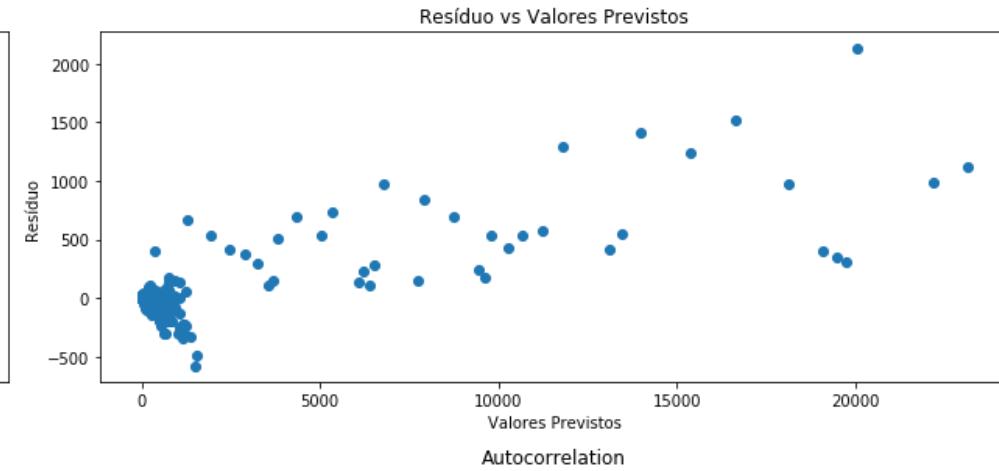
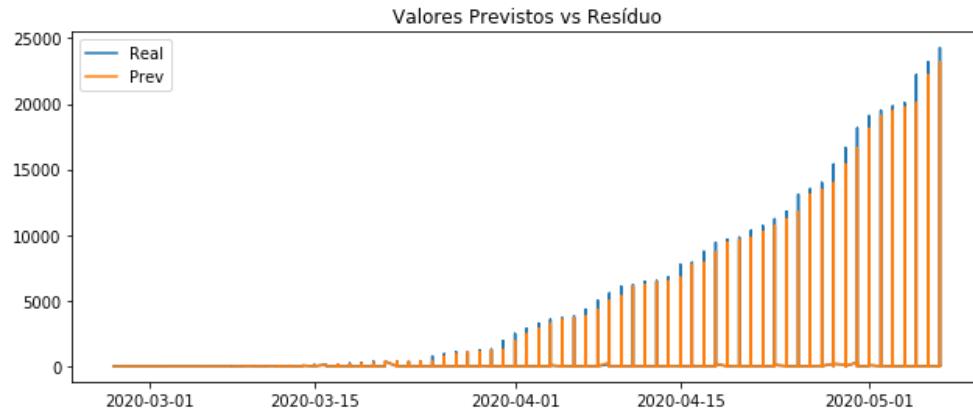
Com a **base teste**, a random forest teve **desempenho muito próximo da baseline, mesmo que ainda não a tenha superado**. O erro médio absoluto foi de 5.6 casos, contra 4.6 casos na baseline (que apenas repete o resultado do dia anterior).

# 6.1. Método de Previsão em Laço

RANDOM FOREST | CASOS ACUMULADOS

152

BASE TESTE – SEM MODIFICAÇÃO DA DISTRIBUIÇÃO DA TARGET



## 6.1. Método de Previsão em Laço

RANDOM FOREST | MORTES ACUMULADAS



153

PREVISÃO  
EM LAÇO

RANDOM  
FOREST

MORTES  
ACUMULADAS



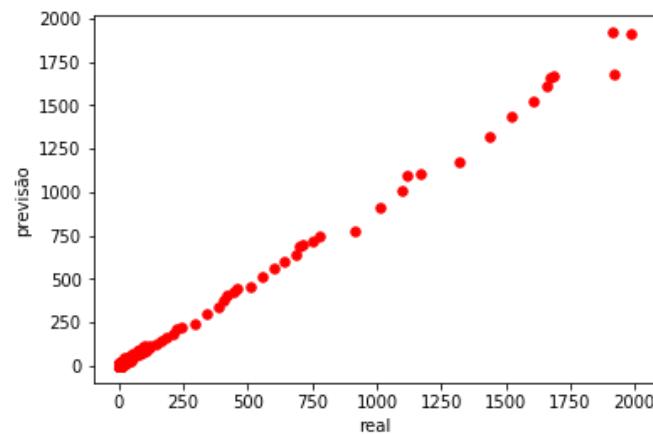
# 6.1. Método de Previsão em Laço

RANDOM FOREST | MORTES ACUMULADAS

154

## BASE TESTE – SEM MODIFICAÇÃO DA DISTRIBUIÇÃO DA TARGET

Random Forest sem modificação da distribuição target:
• Real   Previsto:
• Média: 5.53411   5.36514
• min: 0   0
• 25%: 0   0
• 50%: 0   0
• 75%: 1   1
• max: 1986   1919
• sem previsões negativas



Indicadores para base de teste	BASELINE	Random Forest (laço)
VIÉS	0.36453	0.16898
MSE	21.03945	21.90721
RMSE	4.58688	4.68051
MAE	0.38204	0.48556
MAPE	nan	inf

Com random forest com modificação da distribuição da target, o laço funcionou durante todos os **128 dias**, com indicadores de erros aumentando gradativamente.

Com a **base teste**, a random forest teve **desempenho muito próximo da baseline, mesmo que ainda não a tenha superado**. O erro médio absoluto foi de 0.48 mortes, contra 0.38 mortes na baseline (que apenas repete o resultado do dia anterior).

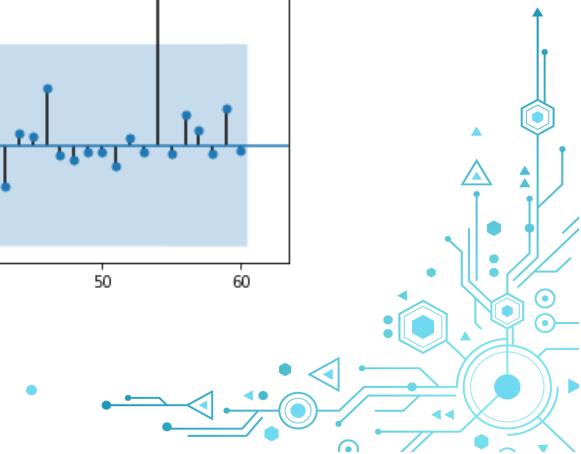
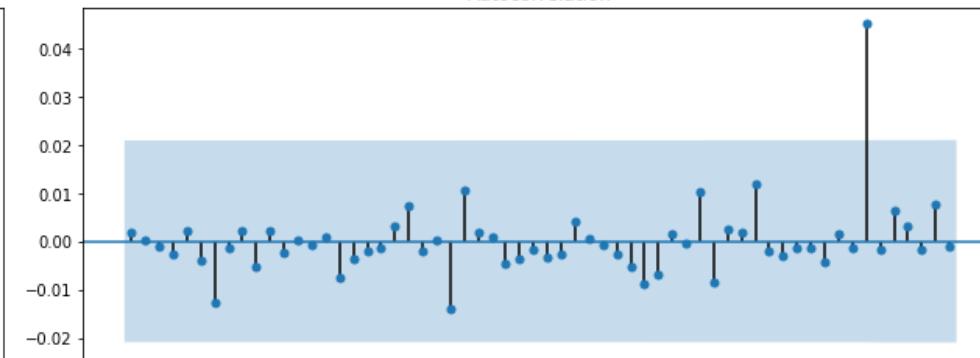
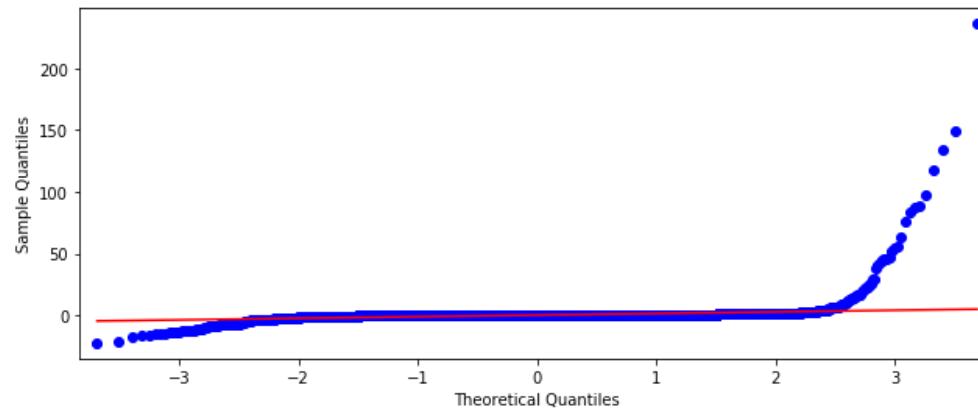
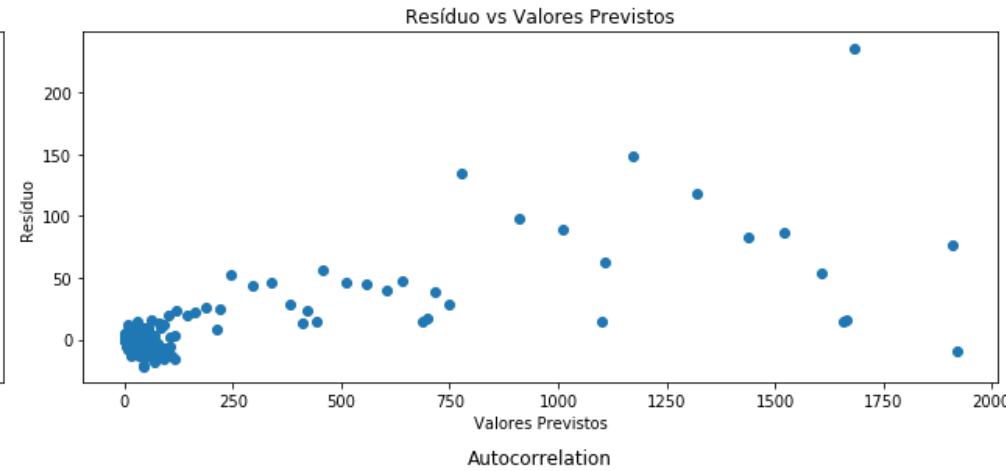
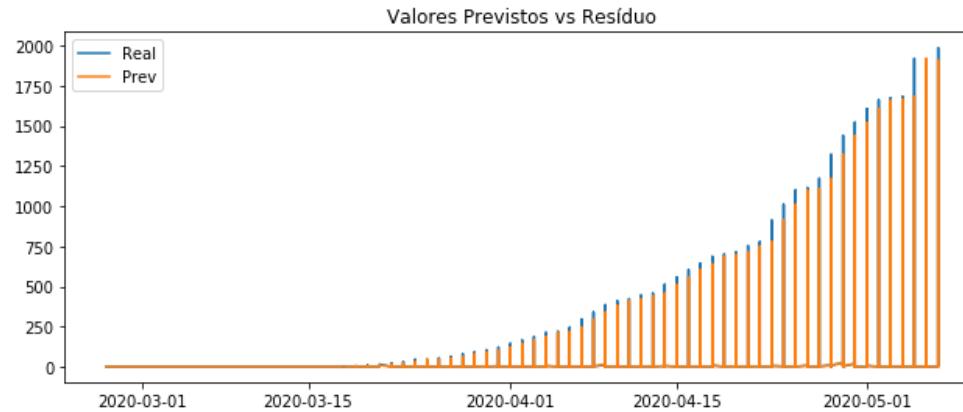


# 6.1. Método de Previsão em Laço

RANDOM FOREST | MORTES ACUMULADAS

155

BASE TESTE – SEM MODIFICAÇÃO DA DISTRIBUIÇÃO DA TARGET



## 6.2. Modelagem com Inteligência Artificial

AUTO MACHINE LEARNING | PYCARET

156

# PYCARET

## 6.2. Modelagem com Inteligência Artificial

AUTO MACHINE LEARNING | PYCARET

157

Como os algoritmos de aprendizado de máquina utilizados não foram suficientes para **superar as métricas da baseline**, optamos pela abordagem com o PyCaret<sup>1</sup>.



### O que é?

- Biblioteca de auto-machine learning para Python

### O que faz?

- Feature Engineering: analisa as variáveis do dataset, cria novas variáveis e seleciona as variáveis mais importantes para o modelo.
- Treina até 24 modelos de regressão e realiza tuning de hiperparâmetros para definir o melhor modelo.

<sup>1</sup> PyCaret.org. PyCaret, April 2020. URL <https://pycaret.org/about>. PyCaret version 1.0.0.



## 6.2. Modelagem com Inteligência Artificial

PYCARET | FEATURE ENGINEERING – CASOS ACUMULADOS

158

Com o Pycaret, criamos uma nova ABT com as seguintes características:

Todas as variáveis anteriores.

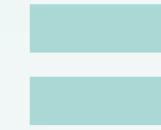


Novas variáveis criadas pelo Pycaret:

- Dummies para variáveis categóricas e de alta cardinalidade.
- Desvio padrão, mínimo, máximo, média, mediana e moda para o conjunto de lags de casos e para o conjunto de lags de mortes.



Normalização dos dados com o método ROBUST



ABT com 155 variáveis



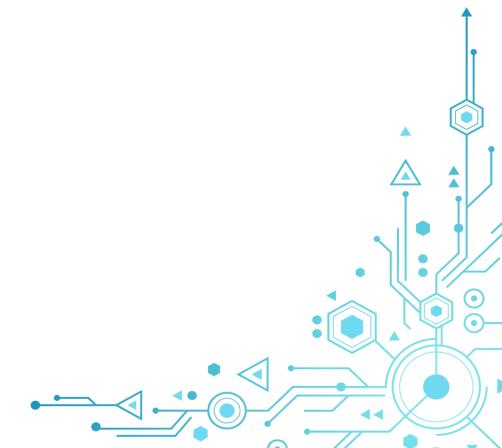
## 6.2. Modelagem com Inteligência Artificial

PYCARET | FEATURE ENGINEERING – CASOS ACUMULADOS

159

### Setup do Pycaret para criação da nova ABT:

Description	Value	Description	Value
0 session_id	14	21 PCA Components	None
1 Transform Target	False	22 Ignore Low Variance	False
2 Transform Target Method	None	23 Combine Rare Levels	False
3 Original Data	(8795, 59)	24 Rare Level Threshold	None
4 Missing Values	False	25 Numeric Binning	False
5 Numeric Features	48	26 Remove Outliers	False
6 Categorical Features	9	27 Outliers Threshold	None
7 Ordinal Features	False	28 Remove Multicollinearity	False
8 High Cardinality Features	True	29 Multicollinearity Threshold	None
9 High Cardinality Method	clustering	30 Clustering	False
10 Sampled Data	(8795, 59)	31 Clustering Iteration	None
11 Transformed Train Set	(7364, 155)	32 Polynomial Features	False
12 Transformed Test Set	(1431, 155)	33 Polynomial Degree	None
13 Numeric Imputer	mean	34 Trigonometry Features	False
14 Categorical Imputer	constant	35 Polynomial Threshold	None
15 Normalize	True	36 Group Features	True
16 Normalize Method	robust	37 Feature Selection	True
17 Transformation	False	38 Features Selection Threshold	0.8
18 Transformation Method	None	39 Feature Interaction	False
19 PCA	False	40 Feature Ratio	False
20 PCA Method	None	41 Interaction Threshold	None



## 6.2. Modelagem com Inteligência Artificial

PYCARET | FEATURE ENGINEERING

160

### ROBUST SCALER

Método de normalização que utiliza estatísticas robustas para outliers.

Remove a mediana de cada entrada e coloca os dados em escala de acordo com o intervalo interquartil (entre o 1º e o 3º quartis).

A centralização e a escala são realizadas de forma independente para cada variável, computando as estatísticas de cada uma.

Como nossos dados possuem muitos outliers, um processo de normalização típico que utiliza a média e a variância tende a perder as nuances, já que os outliers acabam tendo força predominante na média. Por isso o uso da mediana e dos intervalos interquartis tendem a trazer um resultado melhor para o modelo.



## 6.2. Modelagem com Inteligência Artificial

PYCARET | FEATURE ENGINEERING – CASOS ACUMULADOS

161

**ABT gerada pelo Pycaret:**

	mortes_acumuladas_menos13d	lag_de_casos_Median	munuf_12	papel_Interior	Clinicos_Nao_SUS	munuf_42	munuf_36	Obstetrico_Nao_SUS	munuf_29
0	0.00000	-0.16667	0.00000	0.00000	152.05714	0.00000	0.00000	88.75000	0.00000
1	0.00000	-0.16667	0.00000	0.00000	152.05714	0.00000	0.00000	88.75000	0.00000
2	0.00000	-0.16667	0.00000	0.00000	152.05714	0.00000	0.00000	88.75000	0.00000
3	0.00000	-0.16667	0.00000	0.00000	152.05714	0.00000	0.00000	88.75000	0.00000
4	0.00000	-0.16667	0.00000	0.00000	149.97143	0.00000	0.00000	85.66667	0.00000
...	...	...	...	...	...	...	...	...	...
8790	0.00000	0.16667	0.00000	1.00000	-0.25714	0.00000	0.00000	-0.25000	0.00000
8791	0.00000	1.75000	0.00000	1.00000	-0.14286	0.00000	0.00000	0.08333	1.00000
8792	2.00000	1.33333	0.00000	1.00000	-0.25714	0.00000	0.00000	-0.25000	0.00000
8793	1.00000	0.50000	0.00000	1.00000	-0.05714	0.00000	1.00000	-0.08333	0.00000
8794	3.00000	1.83333	0.00000	0.00000	-0.14286	0.00000	0.00000	0.16667	0.00000

8795 rows × 155 columns



## 6.2. Modelagem com Inteligência Artificial

PYCARET | SELEÇÃO DE MODELO – CASOS ACUMULADOS

162

No passo seguinte, o Pycaret testa 24 modelos de regressão até encontrar o que tem a melhor performance nos indicadores de erro. Nesse caso, selecionamos o Theil-Sen Regressor.

	Model	MAE	MSE	RMSE	R2	RMSLE	MAPE	TT (Sec)
0	TheilSen Regressor	3.4295	801.07	27.6972	0.9974	0.2501	0.3128	24.1391
1	Random Sample Consensus	3.8826	1079.8046	31.9443	0.9969	0.3054	0.3851	6.6069
2	Huber Regressor	4.1663	1328.1872	33.9585	0.9955	0.2831	0.3326	1.0151
3	Lasso Regression	6.2854	1426.202	36.5372	0.9937	0.4811	0.7347	0.5298
4	Automatic Relevance Determination	6.4504	1951.6868	42.6671	0.9936	0.4531	0.7786	2.5538
5	Bayesian Ridge	6.5783	1978.5252	42.9785	0.9934	0.4748	0.8685	0.1309
6	Ridge Regression	6.9049	2012.4498	43.3273	0.9932	0.5006	0.9894	0.0186
7	Kernel Ridge	6.9063	2011.527	43.3176	0.9932	0.5001	0.9893	1.9404
8	Elastic Net	6.2132	1687.7628	39.0978	0.9926	0.4582	0.6384	0.4778
9	Orthogonal Matching Pursuit	7.0608	1797.0789	39.8629	0.9924	0.5193	1.2571	0.0209
10	Passive Aggressive Regressor	6.3005	1960.1996	40.6555	0.9906	0.4121	0.9232	0.0584
11	Multi Level Perceptron	8.1006	2365.3889	46.9263	0.988	0.513	1.2768	4.6551
12	Lasso Least Angle Regression	12.8819	7306.8801	81.5128	0.9822	1.0913	3.4211	0.0198
13	CatBoost Regressor	12.1024	13685.0527	100.5753	0.9409	0.3436	1.2036	7.7617
14	K Neighbors Regressor	13.7293	21562.112	134.6666	0.9328	0.458	0.8819	0.2642
15	Gradient Boosting Regressor	13.5654	15956.2671	116.0358	0.9145	0.2682	1.4914	3.1604
16	Extreme Gradient Boosting	11.1135	18451.0445	125.4587	0.8966	0.2179	0.2576	1.0124
17	Random Forest	14.9949	21801.6512	126.6735	0.8636	0.281	1.899	3.2352
18	Support Vector Machine	58.961	521239.607	677.7205	0.0073	0.843	1.1232	7.8711
19	Extra Trees Regressor	28.2692	105835.607	185.6334	0.0054	0.3105	4.8359	2.8888
20	Light Gradient Boosting Machine	45.3278	180377.575	344.1749	-0.2811	0.43	5.7862	0.6016
21	Decision Tree	38.0751	176500.642	272.76	-0.6309	0.3572	6.1207	0.1981
22	AdaBoost Regressor	62.3948	184701.839	295.7479	-0.6744	1.8005	15.0193	0.7501
23	Linear Regression	1654161.55	4.0224E+13	2836384.34	-392888152	2.2106	1108041.94	0.0959
24	Least Angle Regression	1185747011	8.7175E+19	4586268668	-8.4484E+14	13.119	264217151	0.0653

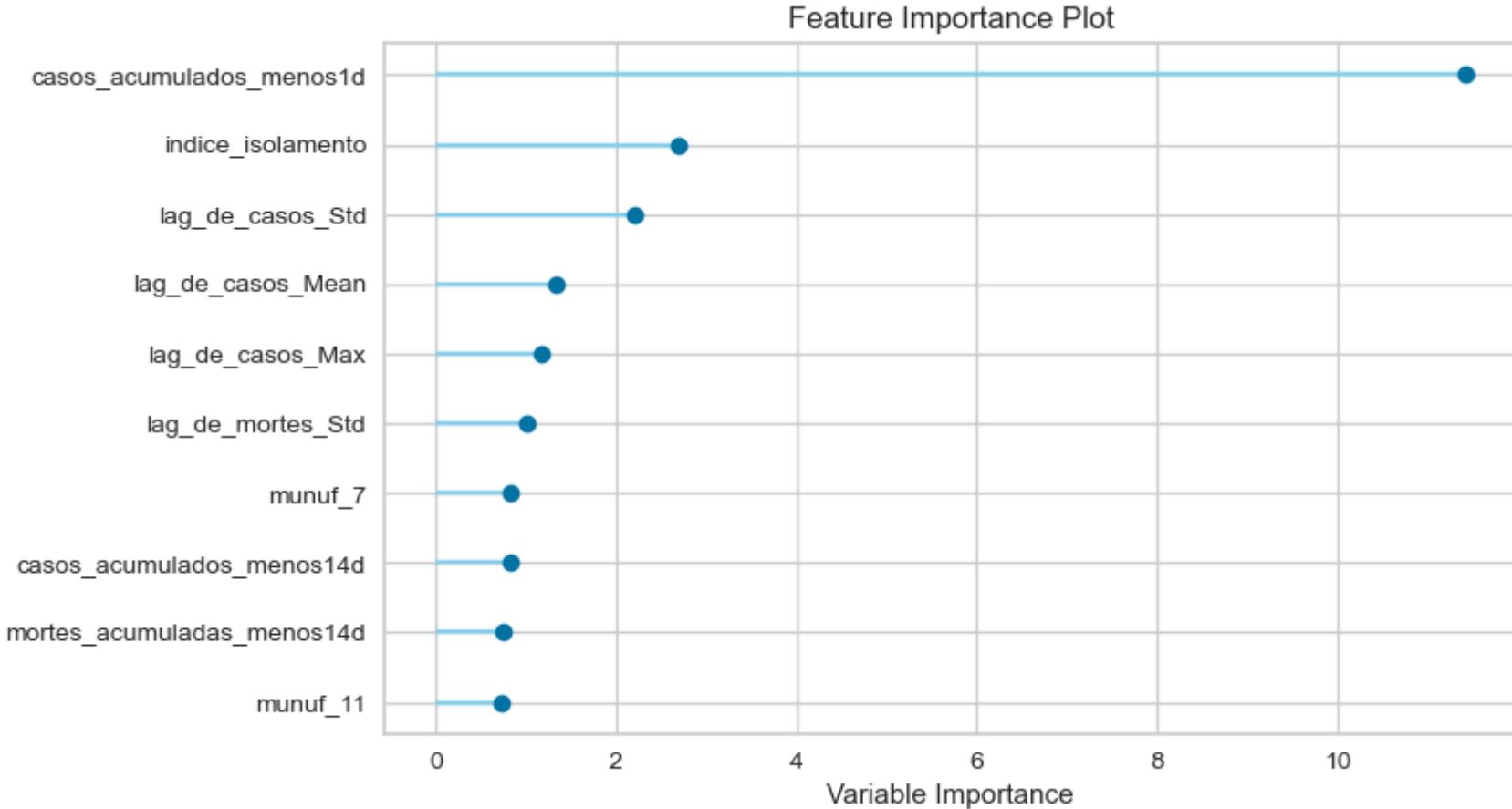


## 6.2. Modelagem com Inteligência Artificial

PYCARET | SELEÇÃO DE MODELO – CASOS ACUMULADOS

163

Após realizar o tuning de parâmetros e encontrar a melhor configuração do Theil-Sen Regressor, verificamos quais são as variáveis mais importantes para o modelo



- A lag de casos de 1 dia se destaca como a mais importante, como o desempenho da baseline já atestara.
- O índice de isolamento aparece como a segunda variável mais importante, confirmando a hipótese sobre sua importância para a curva de contágio.
- Por fim, destacam-se as variáveis criadas pelo Pycaret com as estatísticas para os agrupamentos de lags.



## 6.2. Modelagem com Inteligência Artificial

PYCARET | SELEÇÃO DE MODELO – CASOS ACUMULADOS

164

Configuração  
do modelo final  
para previsão  
de casos  
acumulados:

```
TheilSenRegressor(copy_X=True,  
fit_intercept=True, max_iter=300,  
max_subpopulation=20000,  
n_jobs=-1, n_subsamples=None,  
random_state=14, tol=0.001,  
verbose=False)
```



## 6.2. Modelagem com Inteligência Artificial

PYCARET | FEATURE ENGINEERING – MORTES ACUMULADAS

165

Seguimos o mesmo procedimento para as mortes acumuladas, criando uma nova ABT com o Pycaret, desta vez com 166 variáveis:

Todas as variáveis anteriores.

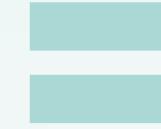


Novas variáveis criadas pelo Pycaret:

- Dummies para variáveis categóricas e de alta cardinalidade.
- Desvio padrão, mínimo, máximo, média, mediana e moda para o conjunto de lags de casos e para o conjunto de lags de mortes.



Normalização dos dados com o método ROBUST



ABT com 166 variáveis



## 6.2. Modelagem com Inteligência Artificial

PYCARET | FEATURE ENGINEERING – MORTES ACUMULADAS

166

### Setup do Pycaret para criação da nova ABT:

Description	Value	Description	Value
0 session_id	14	21 PCA Components	None
1 Transform Target	False	22 Ignore Low Variance	False
2 Transform Target Method	None	23 Combine Rare Levels	False
3 Original Data	(8795, 59)	24 Rare Level Threshold	None
4 Missing Values	False	25 Numeric Binning	False
5 Numeric Features	48	26 Remove Outliers	False
6 Categorical Features	9	27 Outliers Threshold	None
7 Ordinal Features	False	28 Remove Multicollinearity	False
8 High Cardinality Features	True	29 Multicollinearity Threshold	None
9 High Cardinality Method	clustering	30 Clustering	False
10 Sampled Data	(8795, 59)	31 Clustering Iteration	None
11 Transformed Train Set	(7364, 166)	32 Polynomial Features	False
12 Transformed Test Set	(1431, 166)	33 Polynomial Degree	None
13 Numeric Imputer	mean	34 Trigonometry Features	False
14 Categorical Imputer	constant	35 Polynomial Threshold	None
15 Normalize	True	36 Group Features	True
16 Normalize Method	robust	37 Feature Selection	True
17 Transformation	False	38 Features Selection Threshold	0.8
18 Transformation Method	None	39 Feature Interaction	False
19 PCA	False	40 Feature Ratio	False
20 PCA Method	None	41 Interaction Threshold	None



## 6.2. Modelagem com Inteligência Artificial

PYCARET | FEATURE ENGINEERING – MORTES ACUMULADAS

167

**ABT gerada pelo Pycaret:**

	mortes_acumuladas_menos13d	lag_de_casos_Median	munuf_12	papel_Interior	Clinicos_Nao_SUS	munuf_42	munuf_36	Nome_Mesorregiao_13
0	0.00000	-0.16667	0.00000	0.00000	152.05714	0.00000	0.00000	0.00000
1	0.00000	-0.16667	0.00000	0.00000	152.05714	0.00000	0.00000	0.00000
2	0.00000	-0.16667	0.00000	0.00000	152.05714	0.00000	0.00000	0.00000
3	0.00000	-0.16667	0.00000	0.00000	152.05714	0.00000	0.00000	0.00000
4	0.00000	-0.16667	0.00000	0.00000	149.97143	0.00000	0.00000	0.00000
...	...	...	...	...	...	...	...	...
8790	0.00000	0.16667	0.00000	1.00000	-0.25714	0.00000	0.00000	0.00000
8791	0.00000	1.75000	0.00000	1.00000	-0.14286	0.00000	0.00000	0.00000
8792	2.00000	1.33333	0.00000	1.00000	-0.25714	0.00000	0.00000	0.00000
8793	1.00000	0.50000	0.00000	1.00000	-0.05714	0.00000	0.00000	0.00000
8794	3.00000	1.83333	0.00000	0.00000	-0.14286	0.00000	0.00000	0.00000

8795 rows × 166 columns



## 6.2. Modelagem com Inteligência Artificial

PYCARET | SELEÇÃO DE MODELO – MORTES ACUMULADAS

168

Testamos 23 modelos com o Pycaret (tivemos que excluir o modelo de Random Sample Consensus pois ele não funcionou com a base) até chegar ao mesmo resultado: Theil-Sen Regressor.

	Model	MAE	MSE	RMSE	R2	RMSLE	MAPE	TT (Sec)
<b>0</b>	TheilSen Regressor	0.3583	6.725	2.4347	0.997	0.166	0.161	28.8536
<b>1</b>	Huber Regressor	0.4962	16.4902	3.7494	0.9918	0.1688	0.2113	0.9648
<b>2</b>	Bayesian Ridge	0.6357	18.9634	4.0948	0.9892	0.2292	0.3064	0.1322
<b>3</b>	Lasso Regression	0.5631	20.5012	4.1831	0.9889	0.1719	0.2588	0.4459
<b>4</b>	Elastic Net	0.589	21.9725	4.294	0.9879	0.1788	0.288	0.4588
<b>5</b>	Ridge Regression	0.6723	19.5702	4.1726	0.9875	0.2456	0.3205	0.0196
<b>6</b>	Kernel Ridge	0.6722	19.574	4.173	0.9875	0.2454	0.3205	1.9708
<b>7</b>	Automatic Relevance Determination	0.6564	20.1843	4.2385	0.9874	0.2343	0.3111	1.0456
<b>8</b>	Multi Level Perceptron	1.0739	55.6217	7.0518	0.9584	0.3027	0.3853	3.3555
<b>9</b>	Decision Tree	0.9979	116.2211	10.2888	0.9422	0.242	0.1847	0.2413
<b>10</b>	Passive Aggressive Regressor	1.1195	52.3557	6.2105	0.9379	0.3234	0.4572	0.0527
<b>11</b>	K Neighbors Regressor	1.1191	155.3242	11.485	0.9362	0.2533	0.2978	0.3127
<b>12</b>	Orthogonal Matching Pursuit	1.0275	48.3259	5.6793	0.927	0.2979	0.412	0.025
<b>13</b>	Extreme Gradient Boosting	0.8977	90.8444	9.3955	0.9247	0.1753	0.1715	0.9397
<b>14</b>	AdaBoost Regressor	3.3431	150.9928	11.9761	0.9136	1.0592	0.9945	0.9888
<b>15</b>	CatBoost Regressor	1.0168	94.3653	8.8534	0.9106	0.2064	0.2453	7.0651
<b>16</b>	Random Forest	1.0597	97.3791	9.1484	0.8997	0.2374	0.2614	3.8734
<b>17</b>	Gradient Boosting Regressor	1.379	152.7743	10.6182	0.7527	0.234	0.3117	3.1716
<b>18</b>	Extra Trees Regressor	1.3082	151.7709	9.1293	0.7057	0.2551	0.3482	2.6773
<b>19</b>	Support Vector Machine	4.0191	3322.8216	52.3527	0.0665	0.4018	0.4753	5.9779
<b>20</b>	Lasso Least Angle Regression	8.1377	3446.23	53.6723	-0.0058	1.5397	2.3759	0.0241
<b>21</b>	Light Gradient Boosting Machine	3.4705	976.9426	28.1569	-0.1825	0.3583	0.7766	0.5831
<b>22</b>	Linear Regression	8846.3162	3895095396	27914.1995	-9078042.6	1.5056	2299.6244	0.0878
<b>23</b>	Least Angle Regression	4.4409E+24	8.5367E+50	1.3067E+25	-1.4547E+47	27.897	2.3144E+24	0.0792

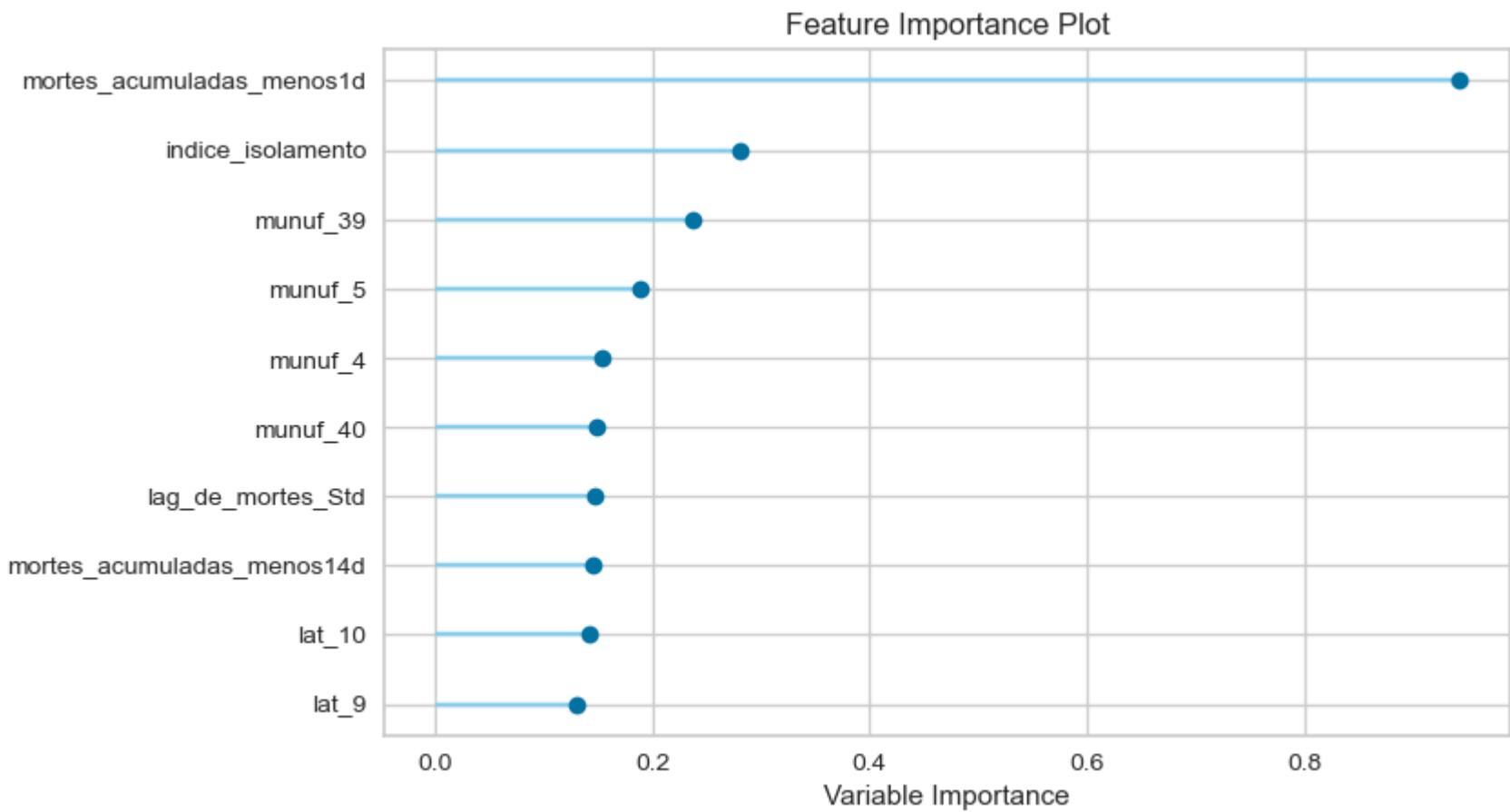


## 6.2. Modelagem com Inteligência Artificial

PYCARET | SELEÇÃO DE MODELO – MORTES ACUMULADAS

169

Após realizar o tuning de parâmetros e encontrar a melhor configuração do Theil-Sen Regressor, verificamos quais são as variáveis mais importantes para o modelo



- A lag de mortes de 1 dia se destaca como a mais importante, como o desempenho da baseline já atestara.
- O índice de isolamento aparece como a segunda variável mais importante, confirmando a hipótese sobre sua importância para a curva de mortes.
- Diferentemente do que aconteceu com os casos, para mortes o município tem mais peso.



## 6.2. Modelagem com Inteligência Artificial

PYCARET | SELEÇÃO DE MODELO – MORTES ACUMULADAS

170

Configuração  
do modelo final  
para previsão  
de mortes  
acumuladas:

```
TheilSenRegressor(copy_X=True,  
fit_intercept=True,  
max_iter=300,  
max_subpopulation=5000,  
n_jobs=-1, n_subsamples=None,  
random_state=14, tol=0.001,  
verbose=False)
```



## 6.2. Modelagem com Inteligência Artificial

PYCARET | REGRESSÃO DE THEIL-SEN

171

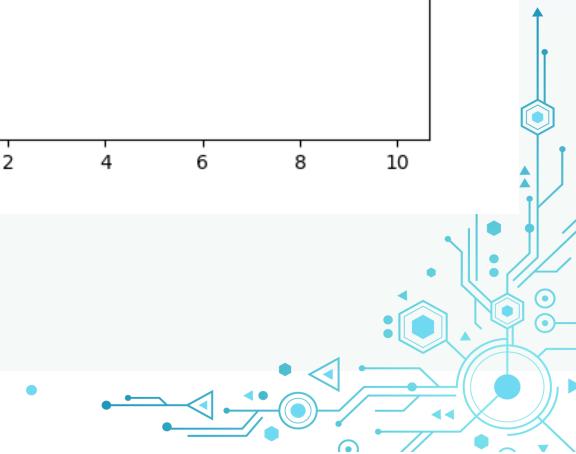
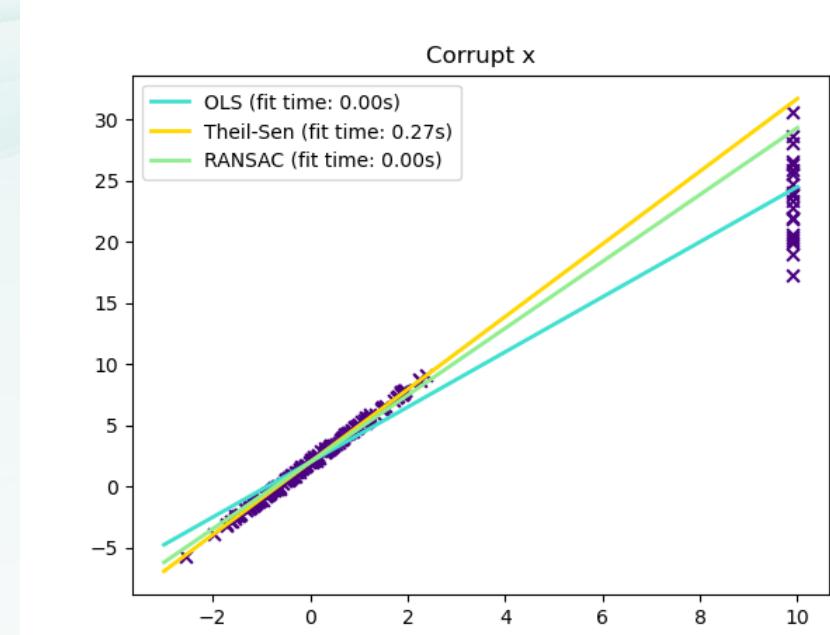
### Regressão de Theil-Sen

Método de regressão desenvolvido por Henri Theil e Pranab K. Sen.

Conforme definido por Theil (1950), o estimador de Theil-Sen de um conjunto de pontos bidimensionais  $(x_i, y_i)$  é a mediana  $M$  das inclinações  $(y_j - y_i) / (x_j - x_i)$  determinada por todos os pares de amostra pontos.

Sen (1968) estendeu essa definição para lidar com o caso em que dois pontos de dados têm a mesma coordenada  $x$ . Na definição de Sen, toma-se a mediana das inclinações definidas apenas a partir de pares de pontos com coordenadas  $x$  distintas.

Como esse método computa a inclinação de todos os pares de pontos com coordenadas diferentes entre si e opta pela mediana das inclinações, acaba sendo mais robusto para conjuntos de dados com outliers.



## 6.2. Modelagem com Inteligência Artificial

COMPARAÇÃO DE MODELOS | REGRESSÃO DE THEIL-SEN

172

Com as ABTs definidas pelo PyCaret e o modelo regressão Theil-Sen, aplicamos a mesma metodologia definida anteriormente: o teste com a previsão em laço e o teste com previsão tradicional, para definir o melhor modelo.



Durante os testes, a regressão de Theil-Sen eventualmente realizou algumas previsões negativas. Então, aplicamos um script com a regra de que sempre que a previsão da regressão Theil-Sen for negativa, o resultado deve ser trocado pela baseline, ou seja, a lag de 1 dia.



## 6.2. Modelagem com Inteligência Artificial

COMPARAÇÃO DE MODELOS | RESULTADOS

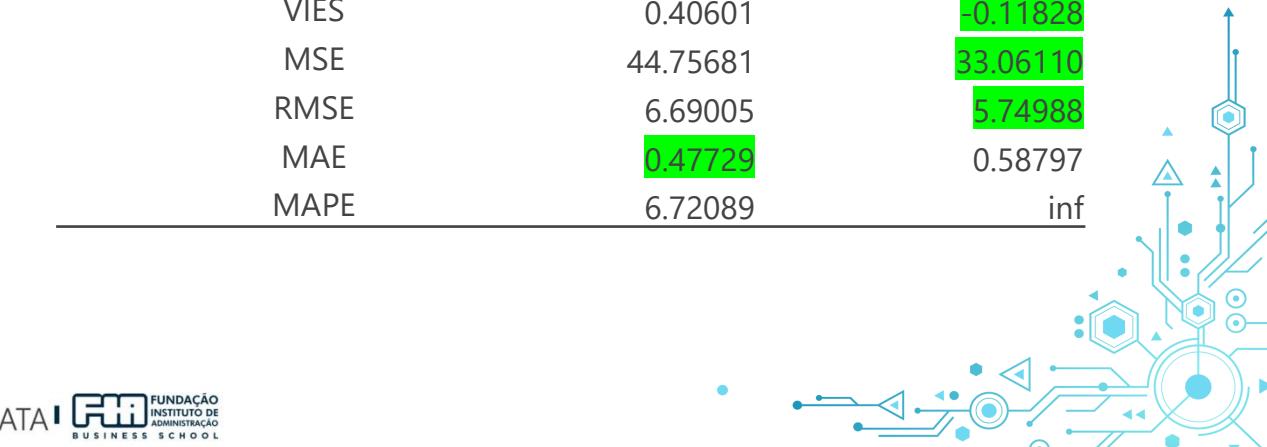
173

Base Teste - Casos Acumulados Método de Previsão em Laço	Baseline	Theil-Sen Regressor + Baseline para Previsões Negativas
VIÉS	4.54008	-0.84182
MSE	2534.639	1470.94326
RMSE	50.3452	38.35288
MAE	4.6581	4.17728
MAPE	11.77102	19.92572

Base Teste - Casos Acumulados Método de Previsão Tradicional	Baseline	Theil-Sen Regressor + Baseline para Previsões Negativas
VIÉS	5.69602	0.02320
MSE	4835.93082	1532.50345
RMSE	69.54086	39.14720
MAE	5.92383	4.29398
MAPE	11.07093	15.31185

Base Teste – Mortes Acumuladas Método de Previsão em Laço	Baseline	Theil-Sen Regressor + Baseline para Previsões Negativas
VIÉS	0.36453	-0.05720
MSE	21.03945	12.82113
RMSE	4.58688	3.58066
MAE	0.38204	0.37105
MAPE	nan	inf

Base Teste – Mortes Acumuladas Método de Previsão Tradicional	Baseline	Theil-Sen Regressor + Baseline para Previsões Negativas
VIÉS	0.40601	-0.11828
MSE	44.75681	33.06110
RMSE	6.69005	5.74988
MAE	0.47729	0.58797
MAPE	6.72089	inf



## 6.2. Modelagem com Inteligência Artificial

COMPARAÇÃO DE MODELOS | RESULTADOS

174

### Conclusões:

A regressão de Theil-Sen aplicada sobre a ABT com normalização robusta conseguiu **superar** a baseline.

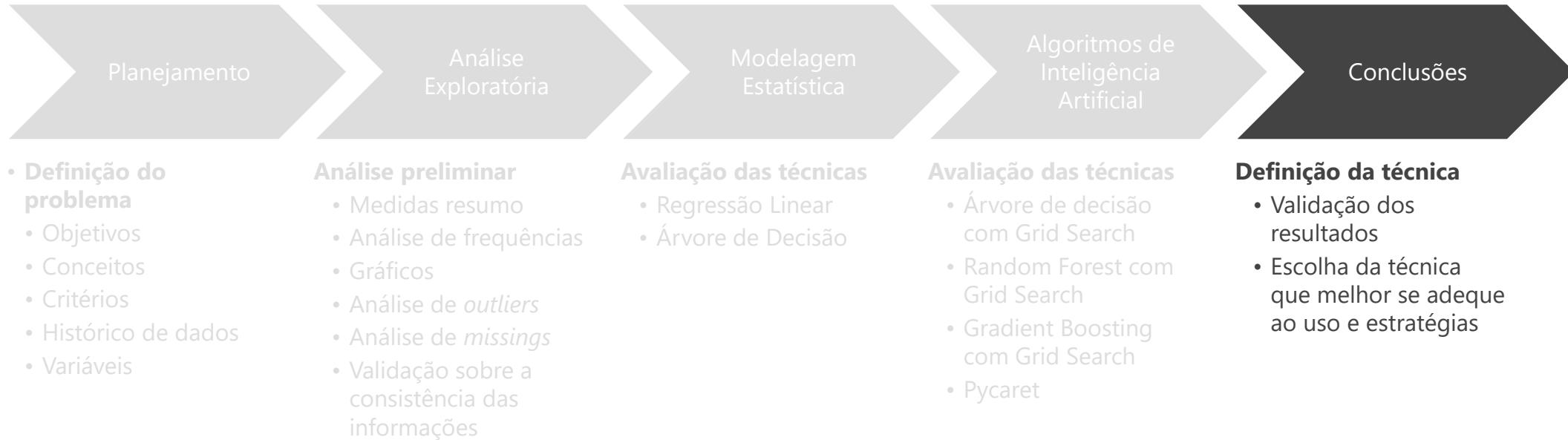
O método de **previsão em laço** teve performance melhor do que o método tradicional, pois ele é treinado com uma base maior a cada iteração.

Para colocar em produção, é fundamental que o modelo de série temporal seja constantemente retreinado.



# Metodologia de Análise de Dados

175



# 7. Conclusões

176

A **principal dificuldade** no trabalho de previsão de casos de COVID-19 e mortes por COVID-19 é poder **contar com uma base de dados sólida**.

Nesse sentido, o **governo** (principalmente o federal, que deveria ter papel de liderança, mas na sua ausência os governos estaduais e municipais) tem a **oportunidade de definir padrões e parâmetros** para que os entes da federação façam o **registro dos casos e mortes** da forma mais fidedigna possível. Se a informação fosse tratada de forma séria, seria um primeiro passo para tratar a pandemia de forma séria.

O trabalho com **séries temporais** **funciona melhor com o método de previsão em laço**, que pressupõe que a cada dia o modelo será retreinado com todos os dados dos dias anteriores e que a previsão será realizada apenas para o dia atual, **aumentando sua precisão**. Quanto menor a quantidade de dados e quanto mais ao longo do tempo se tenta prever, menor é a precisão da previsão.

O **PyCaret** se mostra uma **potente ferramenta** para dar insights no **feature engineering e na definição do melhor modelo**, diminuindo drasticamente o tempo de preparação da modelagem de dados. Recomendamos o uso em produção em etapas de revalidação do modelo, quando seus indicadores de erro forem superados pela baseline.

O **melhor modelo** para previsão de casos de COVID-19 e mortes por COVID-19 no estado de São Paulo é a **regressão de Theil-Sen** aplicada sobre **ABT com normalização robusta**, devido ao grande número de outliers, somado a **um script que troque previsões negativas pelo resultado da lag de 1 dia**. Com isso, os modelos superam a baseline, reduzindo o MAE de casos de 4.6581 para 4.17728 e o MAE de mortes de 0.38204 para 0.37105.

As variáveis mais importantes para previsão de casos de COVID-19 são: **lag de casos de 1 dia, índice de isolamento**, desvio padrão, média e máximo do conjunto de lags de casos de 1 a 14 dias. As variáveis mais importantes para previsão de mortes por COVID-19 são: **lag de mortes de 1 dia, índice de isolamento**, município.

O fato das lags de 1 dia se mostrarem as variáveis mais importantes para os modelos de regressão de Theil-Sen confirma o que foi visto com a baseline, que usa como previsão para o dia atual justamente o resultado do dia anterior, e que foi tão difícil de superar.

**Aplicação prática** do modelo de previsão diária: realizando o cálculo da previsão acumulada menos o acumulado do dia anterior, o resultado é o número de casos e mortes do dia. Com isso, os **governos municipais podem se preparar logisticamente** dia a dia para os novos casos e mortes previstos (reservando ou remanejando leitos de UTI, por exemplo).



## 7. Conclusões – Validação dos Resultados

APLICAÇÃO DE MODELO | REGRESSÃO DE THEIL SEN

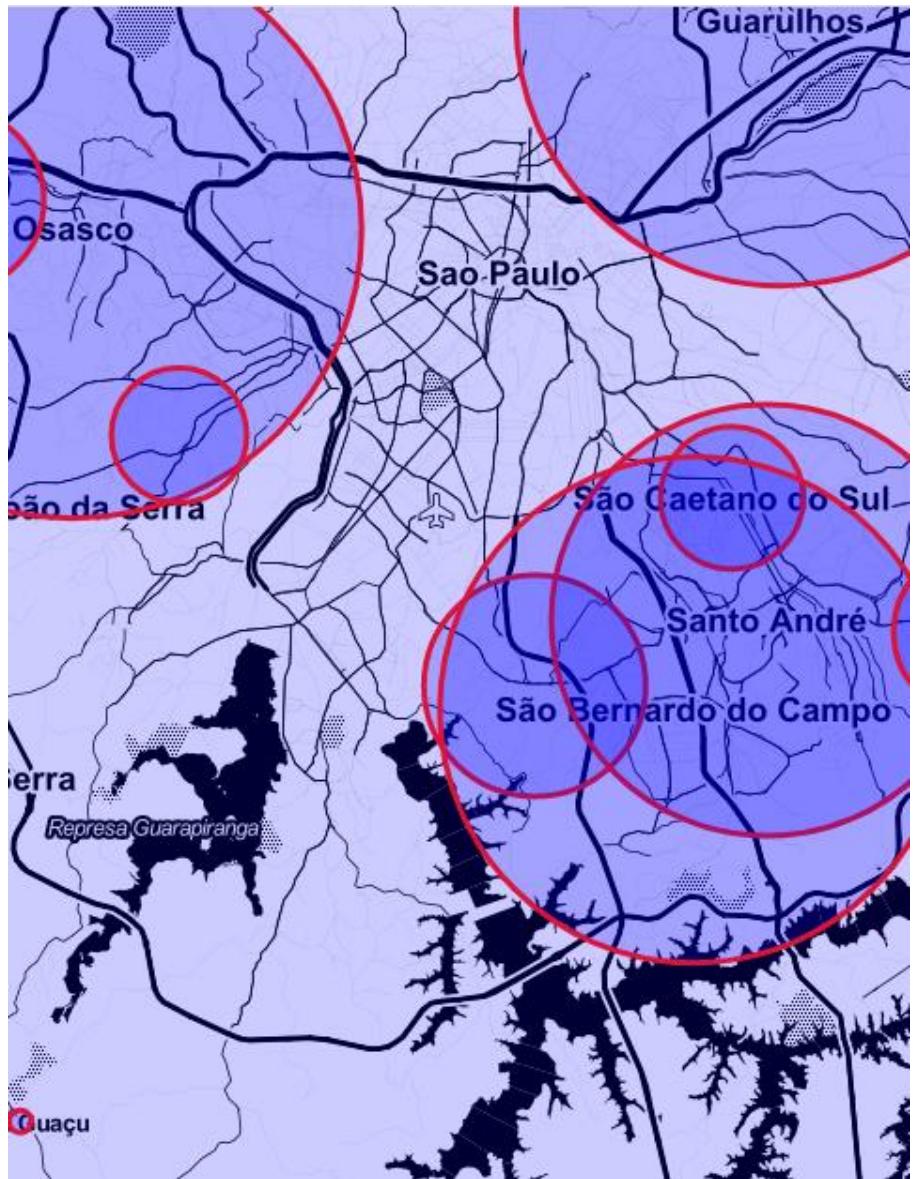
177

# REGRESSÃO DE THEIL-SEN

# 7.1. Método de Previsão em Laço

REGRESSÃO DE THEIL-SEN | CASOS ACUMULADOS

178



PREVISÃO  
EM LAÇO

REGRESSÃO  
DE THEIL-  
SEN

CASOS  
ACUMULADOS



### 7.1.1. Método de Previsão em Laço

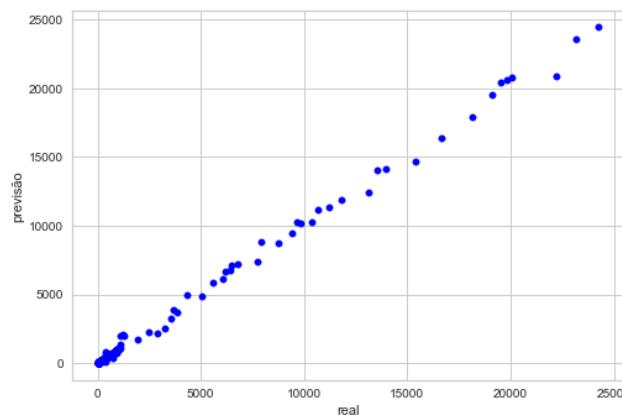
## REGRESSÃO DE THEIL-SEN | CASOS ACUMULADOS

# BASE TESTE – MUNICÍPIOS DO ESTADO DE SP

Indicadores para base de teste	BASELINE	Theil-Sen Regressor + Baseline para Previsões Negativas (laço)
VIÉS	4.54008	-0.84182
MSE	2534.639	1470.94326
RMSE	50.3452	38.35288
MAE	4.6581	4.17728
MAPE	11.77102	19.92572

Com a ABT com normalização robusta e a regressão de Theil-Sen, o laço funcionou durante todos os **128 dias**.

Com a **base teste**, a regressão de Theil-Sen em conjunto com o script de substituição de negativos pela lag de 1 dia teve **desempenho superior à baseline**. O erro médio absoluto foi de 4.17 casos, contra 4.65 casos na baseline (que apenas repete o resultado do dia anterior).

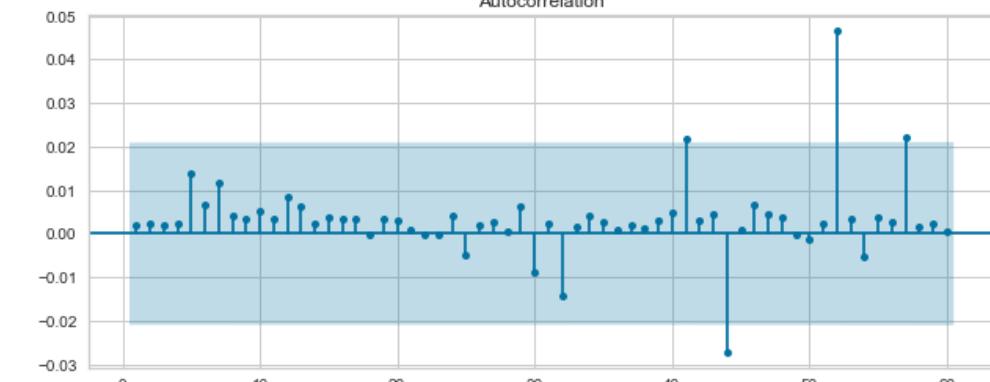
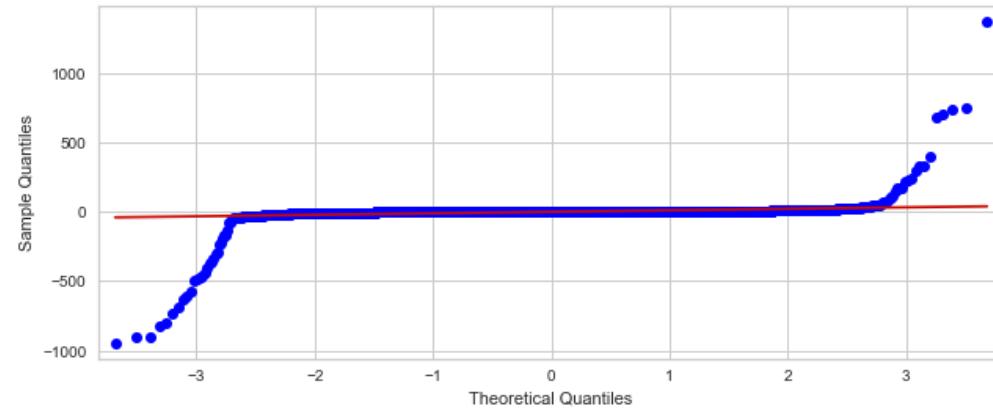
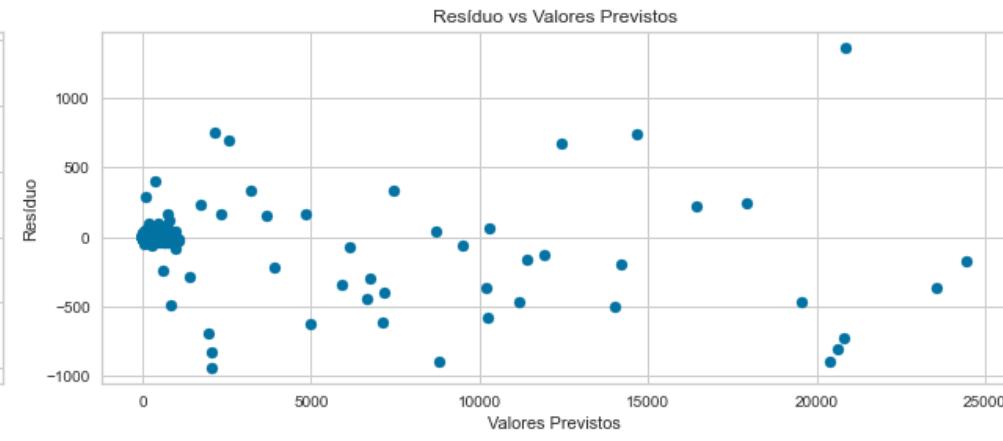
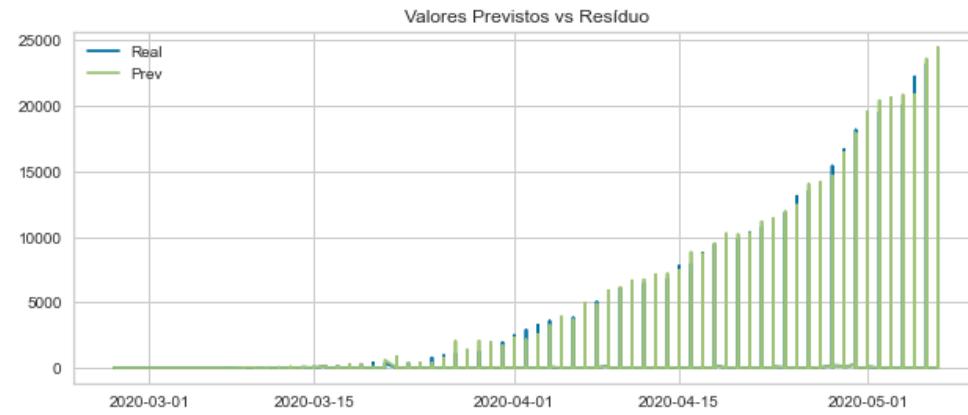


## 7.1.1. Método de Previsão em Laço

REGRESSÃO DE THEIL-SEN | CASOS ACUMULADOS

180

BASE TESTE – MUNICÍPIOS DO ESTADO DE SP

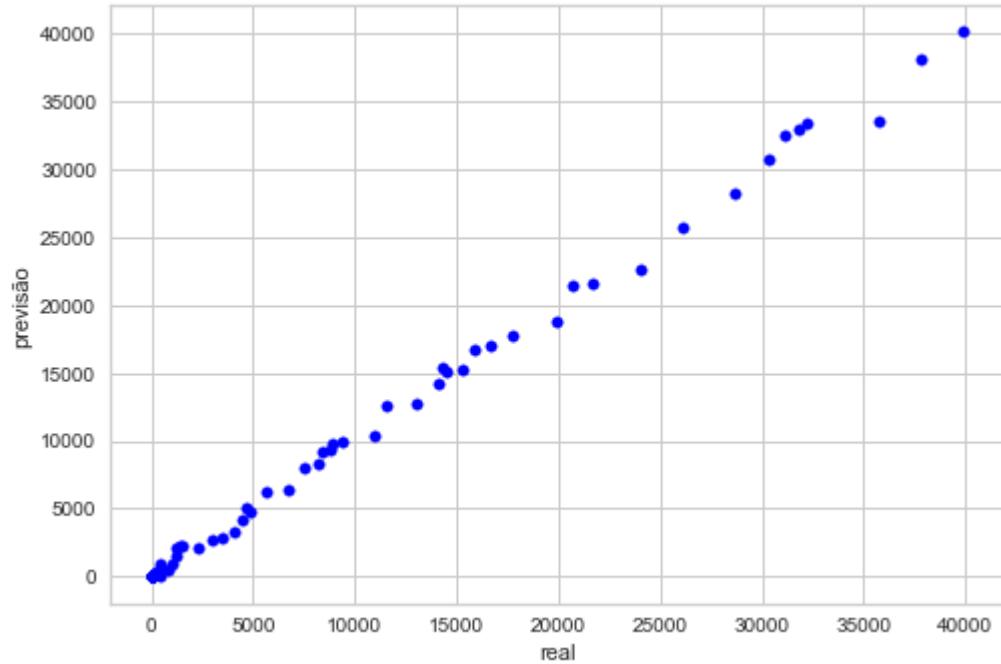


## 7.1.1. Método de Previsão em Laço

REGRESSÃO DE THEIL-SEN | CASOS ACUMULADOS

181

### BASE TESTE – ESTADO DE SP



Indicadores para base de teste	BASELINE	Theil-Sen Regressor + Baseline para Previsões Negativas (laço)
VIÉS	562.38028	-104.26761
MSE	887629.73239	347980.04225
RMSE	942.14104	589.89833
MAE	562.38028	392.40845
MAPE	12.30630	17.46522

Com a **base teste**, a regressão de Theil-Sen em conjunto com o script de substituição de negativos pela lag de 1 dia teve **desempenho superior à baseline**. O erro médio absoluto foi de 392 casos, contra 562 casos na baseline (que apenas repete o resultado do dia anterior).

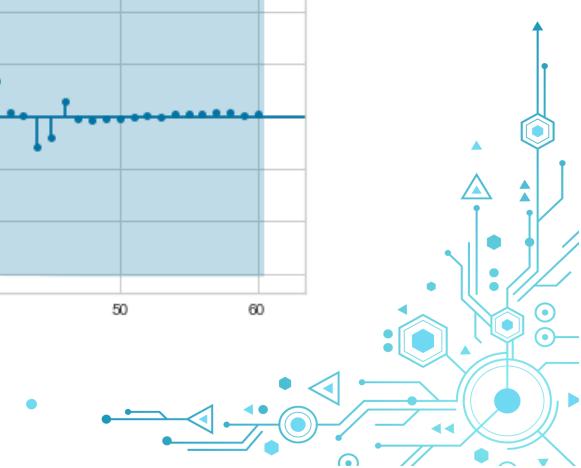
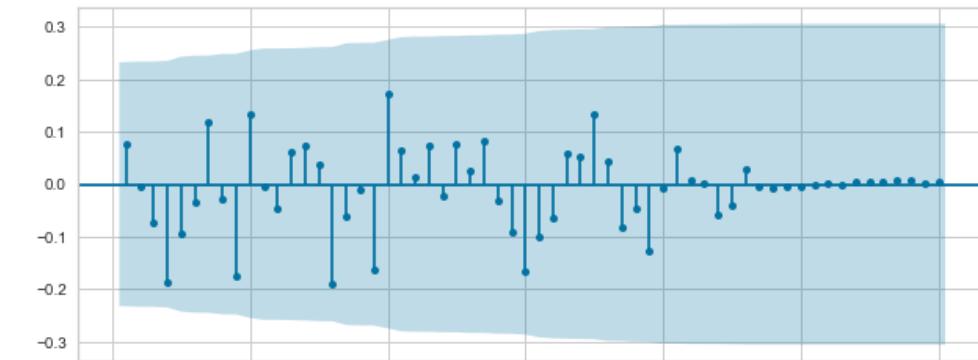
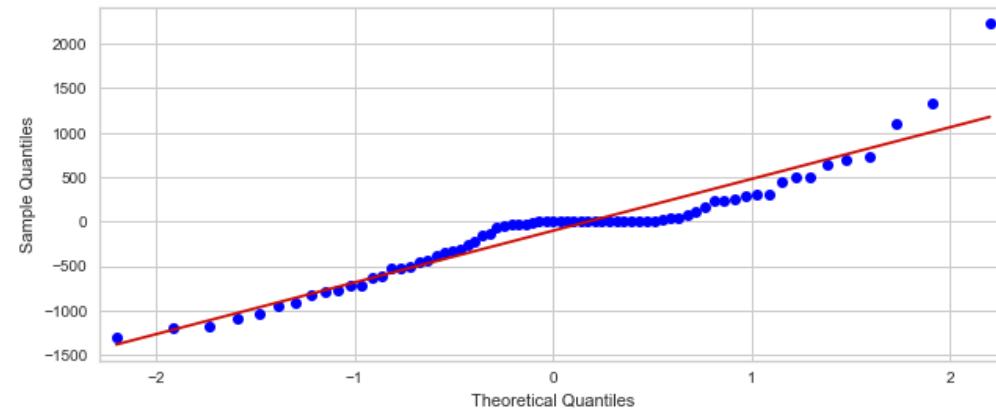
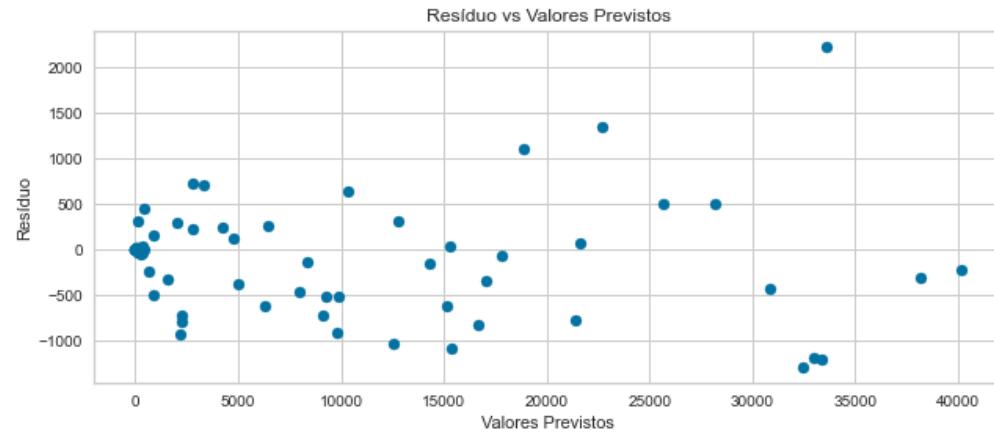
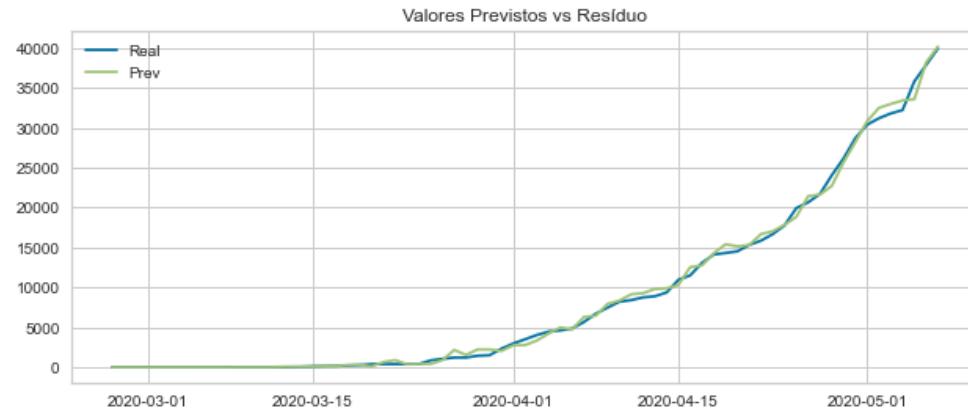


## 7.1.1. Método de Previsão em Laço

REGRESSÃO DE THEIL-SEN | CASOS ACUMULADOS

182

BASE TESTE – ESTADO DE SP

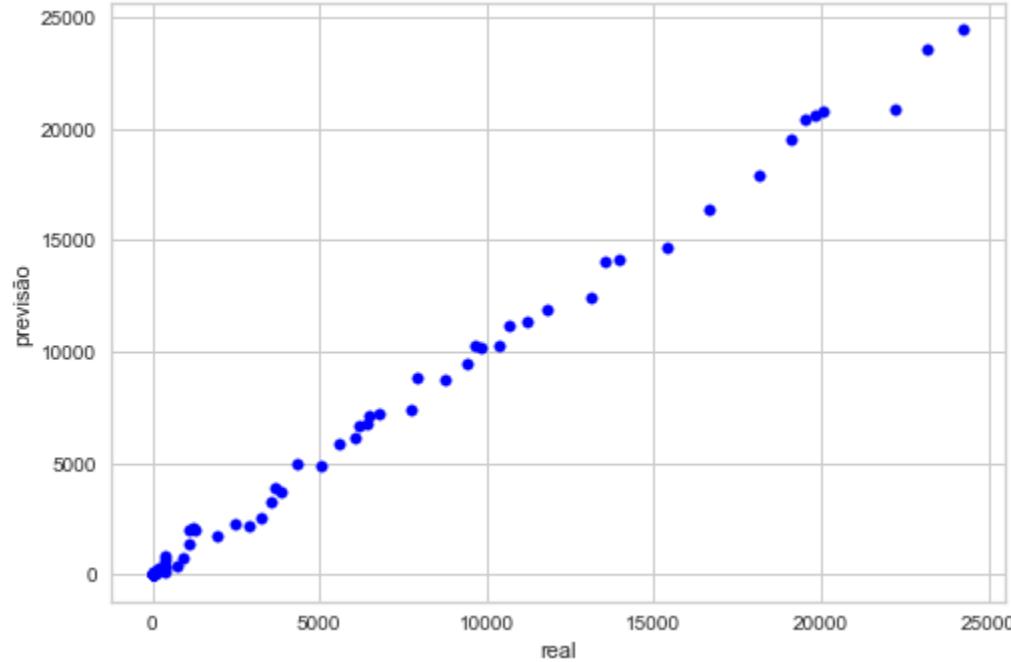


## 7.1.1. Método de Previsão em Laço

REGRESSÃO DE THEIL-SEN | CASOS ACUMULADOS

183

BASE TESTE – MUNICÍPIO DE SÃO PAULO



Indicadores para base de teste	BASELINE	Theil-Sen Regressor + Baseline para Previsões Negativas (laço)
VIÉS	341.85915	-88.85915
MSE	308363.77465	179010.63380
RMSE	555.30512	423.09648
MAE	341.85915	289.98592
MAPE	11.73664	18.38914

Com a **base teste**, a regressão de Theil-Sen em conjunto com o script de substituição de negativos pela lag de 1 dia teve **desempenho superior à baseline**. O erro médio absoluto foi de 289 casos, contra 341 casos na baseline (que apenas repete o resultado do dia anterior).

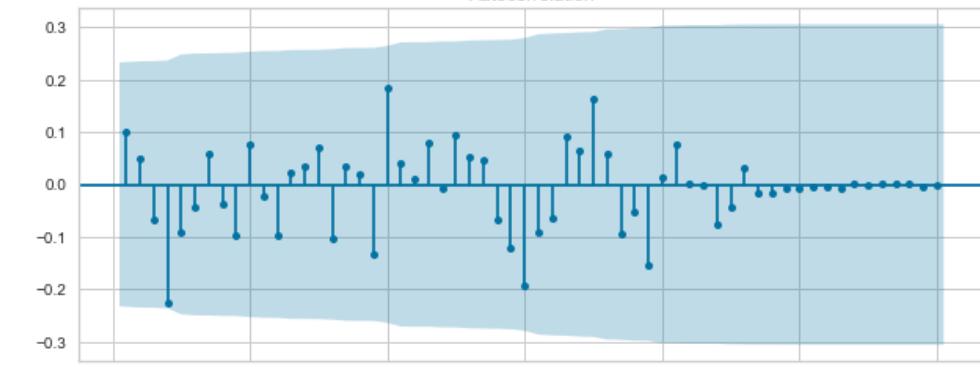
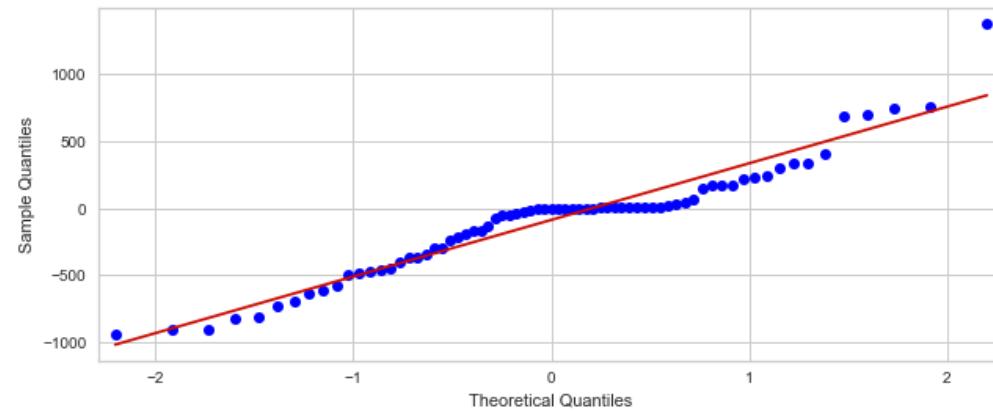
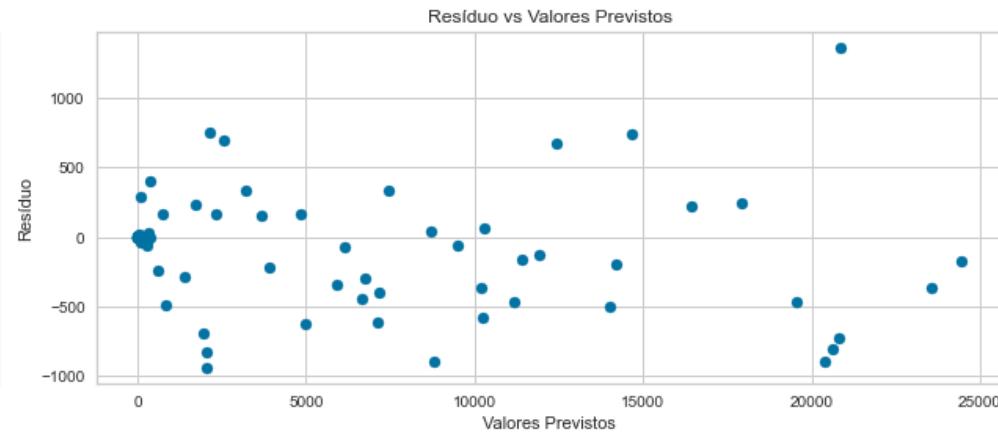
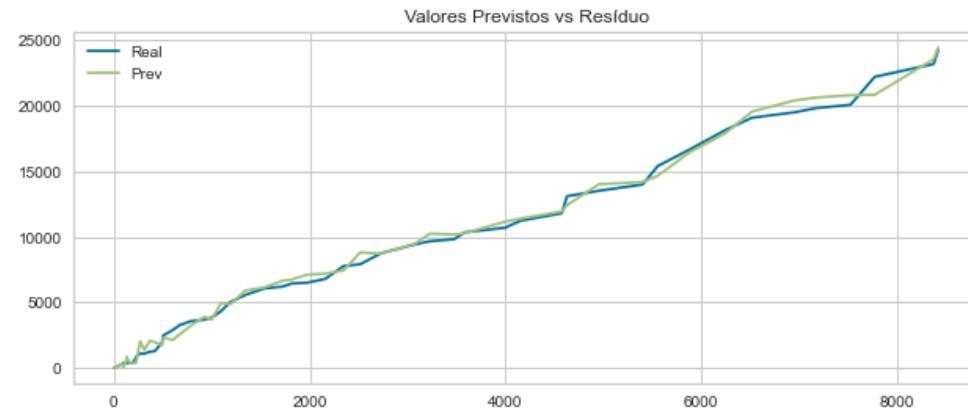


## 7.1.1. Método de Previsão em Laço

REGRESSÃO DE THEIL-SEN | CASOS ACUMULADOS

184

BASE TESTE – MUNICÍPIO DE SÃO PAULO

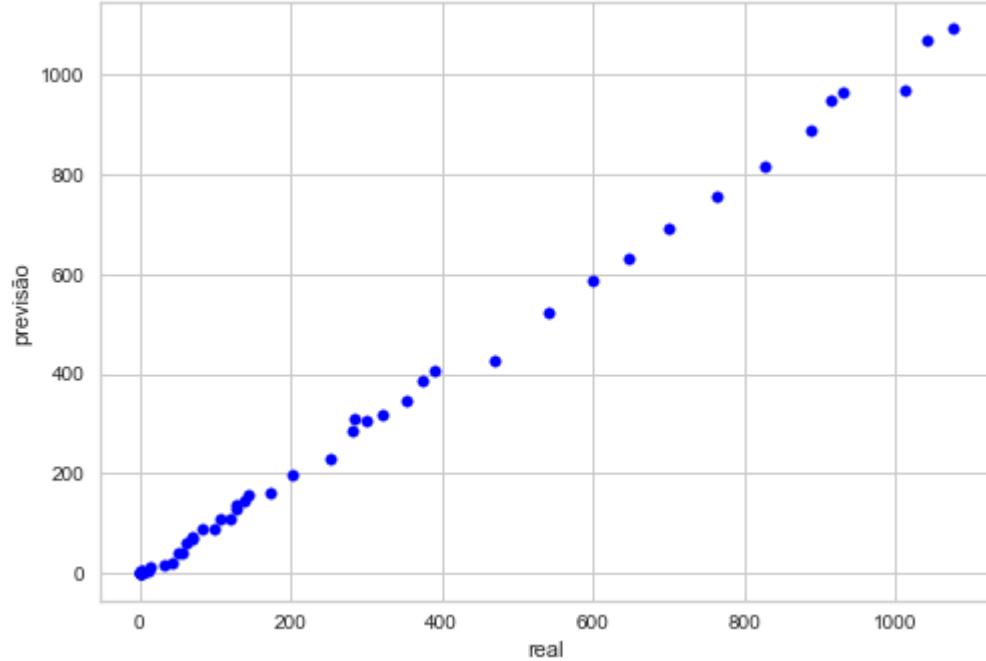


## 7.1.1. Método de Previsão em Laço

REGRESSÃO DE THEIL-SEN | CASOS ACUMULADOS

185

### BASE TESTE – MUNICÍPIO DE OSASCO



Indicadores para base de teste	BASELINE	Theil-Sen Regressor + Baseline para Previsões Negativas (laço)
VIÉS	21.50000	1.76000
MSE	1013.30000	240.60000
RMSE	31.83237	15.51129
MAE	21.50000	10.80000
MAPE	13.13095	22.10773

Com a **base teste**, a regressão de Theil-Sen em conjunto com o script de substituição de negativos pela lag de 1 dia teve **desempenho superior à baseline**. O erro médio absoluto foi de 10.8 casos, contra 21.5 casos na baseline (que apenas repete o resultado do dia anterior).

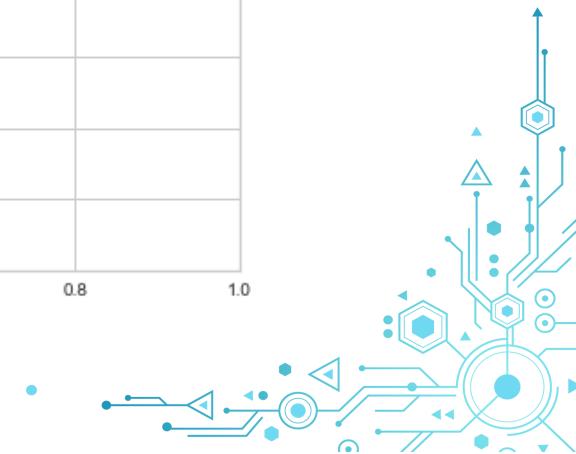
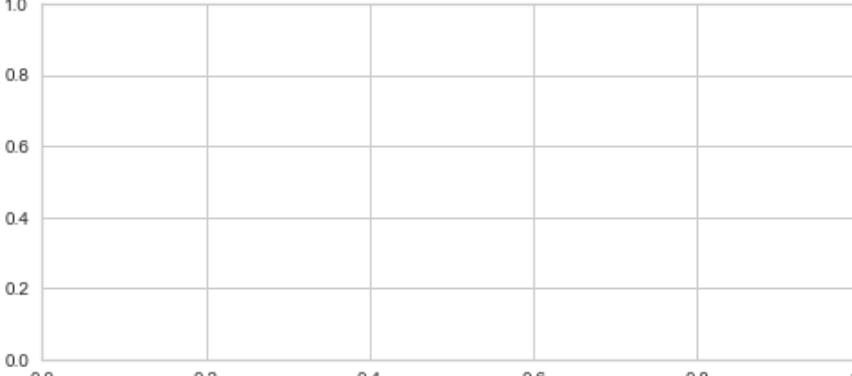
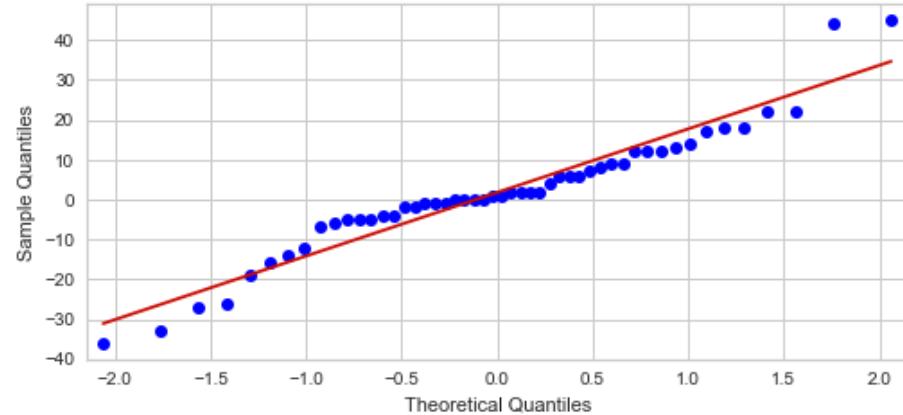
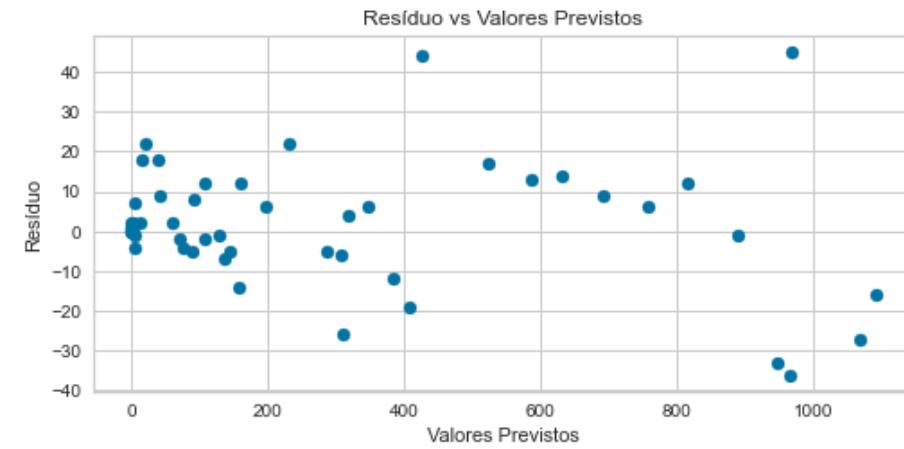
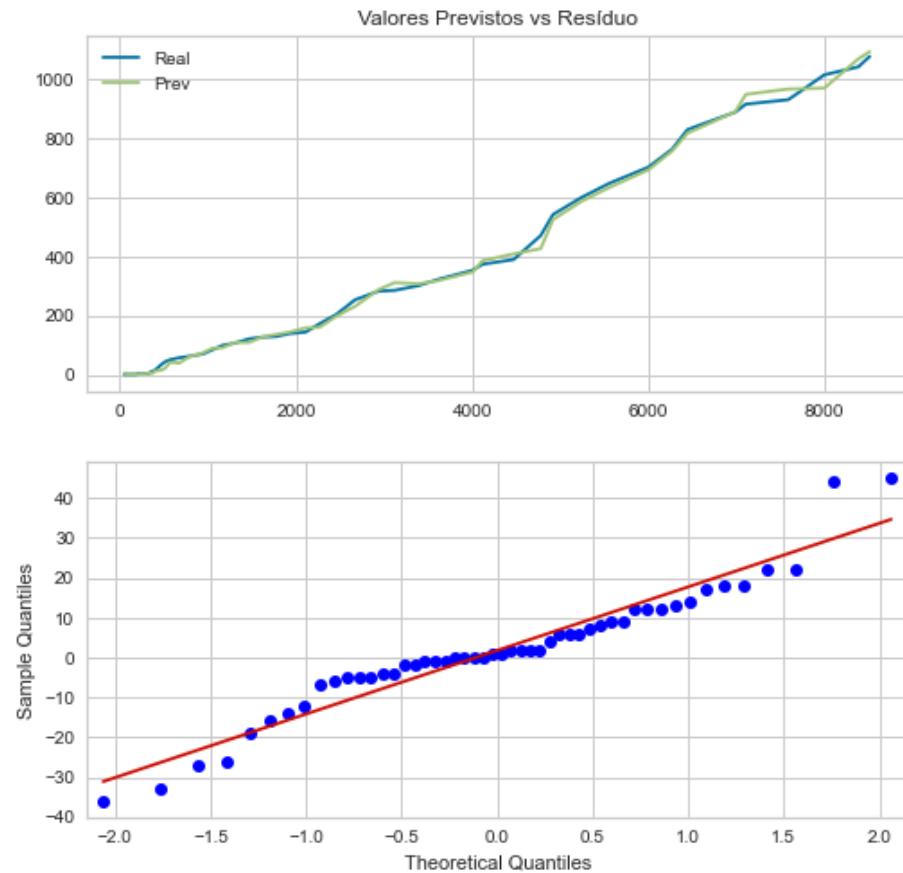


## 7.1.1. Método de Previsão em Laço

REGRESSÃO DE THEIL-SEN | CASOS ACUMULADOS

186

BASE TESTE – MUNICÍPIO DE OSASCO

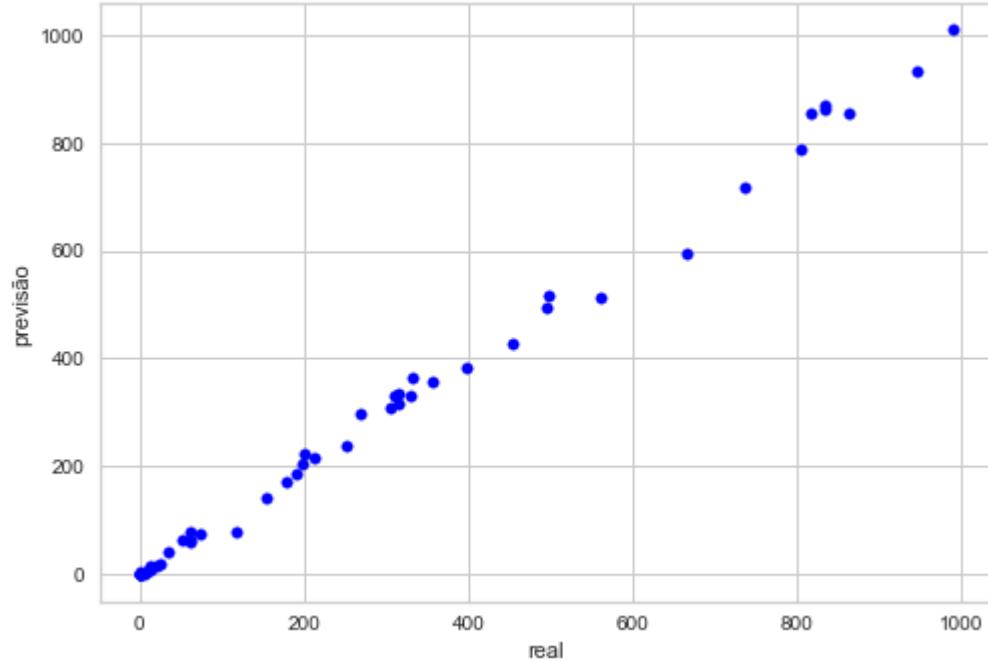


## 7.1.1. Método de Previsão em Laço

REGRESSÃO DE THEIL-SEN | CASOS ACUMULADOS

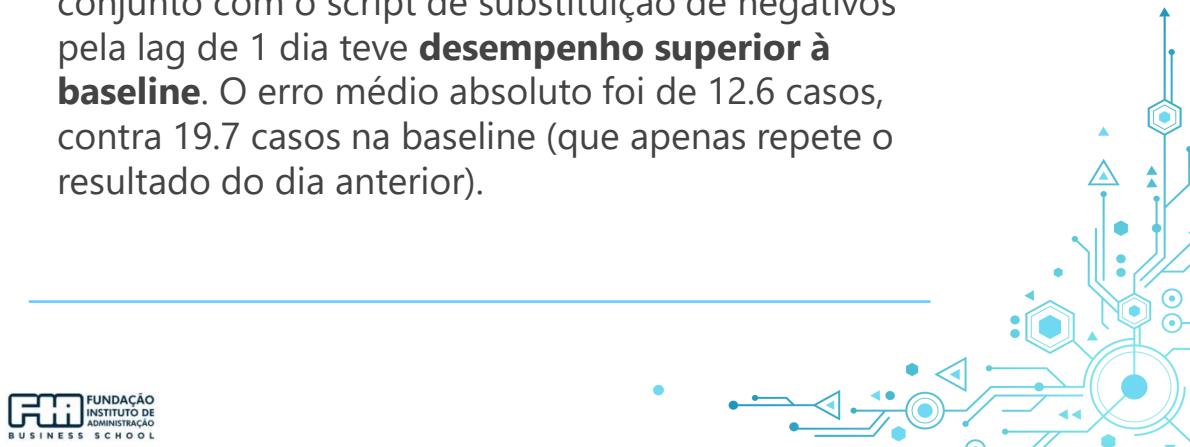
187

BASE TESTE – MUNICÍPIO DE GUARULHOS



Indicadores para base de teste	BASELINE	Theil-Sen Regressor + Baseline para Previsões Negativas (laço)
VIÉS	19.01923	0.01923
MSE	960.55769	376.59615
RMSE	30.99287	19.40609
MAE	19.17308	12.67308
MAPE	12.40535	18.25936

Com a **base teste**, a regressão de Theil-Sen em conjunto com o script de substituição de negativos pela lag de 1 dia teve **desempenho superior à baseline**. O erro médio absoluto foi de 12.6 casos, contra 19.7 casos na baseline (que apenas repete o resultado do dia anterior).

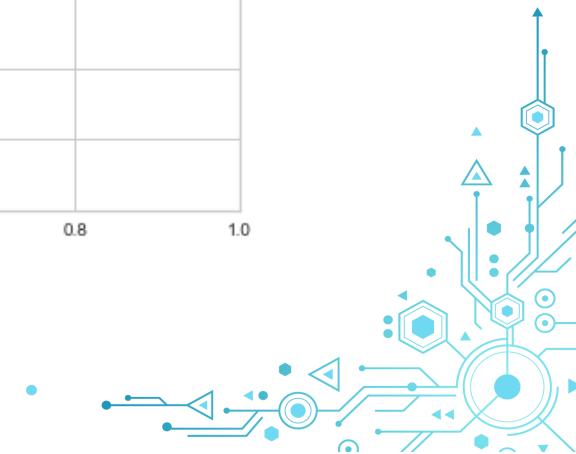
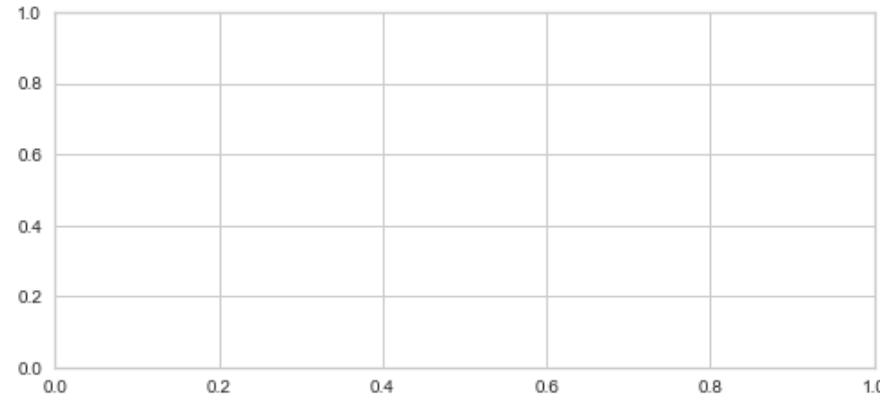
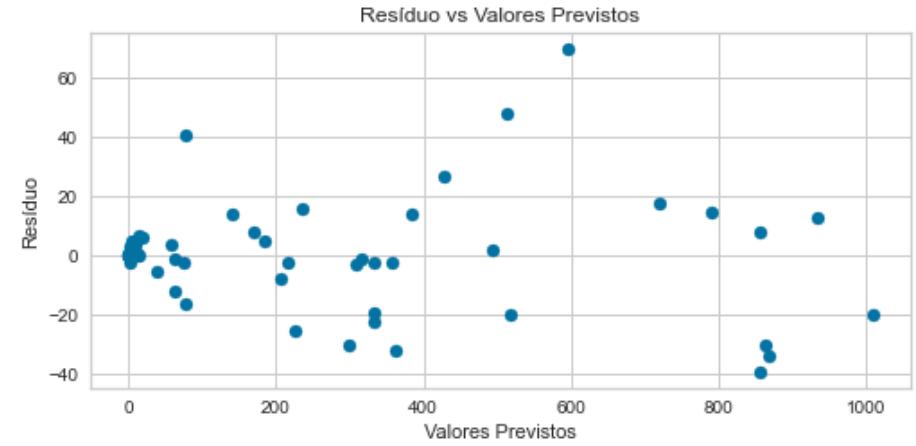
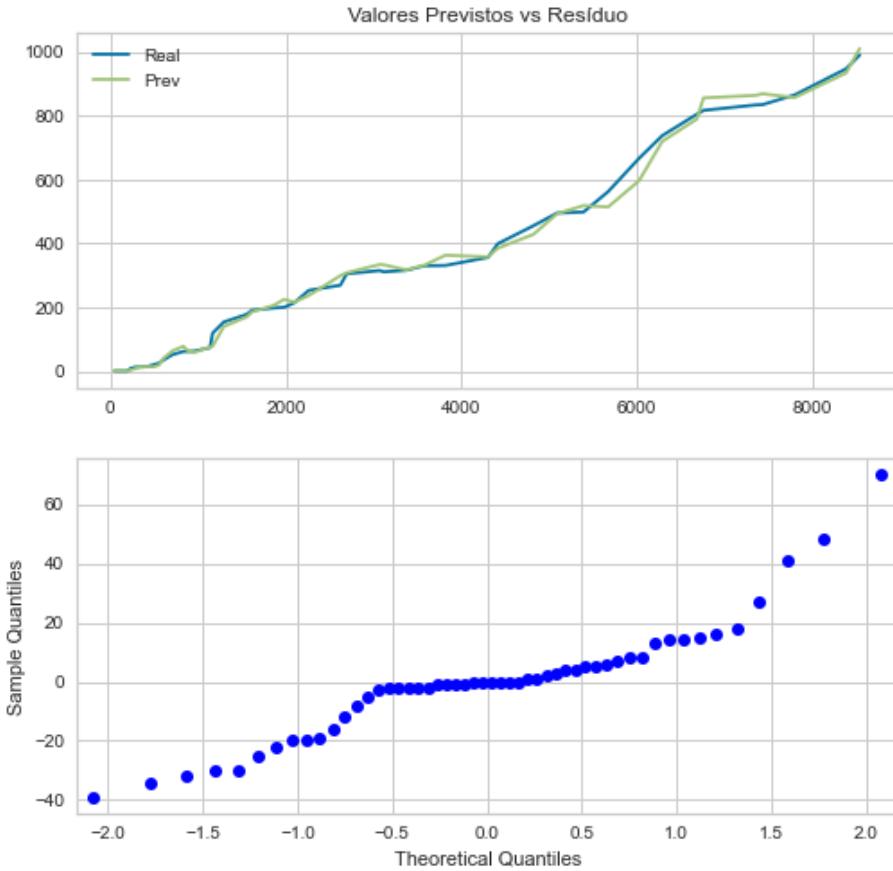


## 7.1.1. Método de Previsão em Laço

REGRESSÃO DE THEIL-SEN | CASOS ACUMULADOS

188

BASE TESTE – MUNICÍPIO DE GUARULHOS

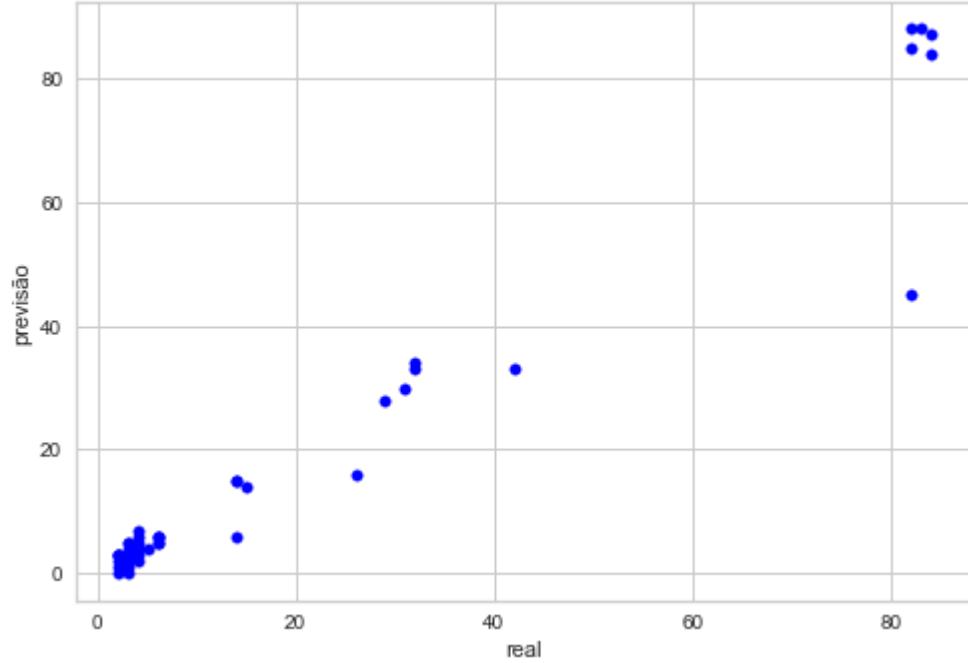


## 7.1.1. Método de Previsão em Laço

REGRESSÃO DE THEIL-SEN | CASOS ACUMULADOS

189

BASE TESTE – MUNICÍPIO DE SÃO SEBASTIÃO



Indicadores para base de teste	BASELINE	Theil-Sen Regressor + Baseline para Previsões Negativas (laço)
VIÉS	1.88636	1.04545
MSE	43.47727	39.95455
RMSE	6.59373	6.32096
MAE	1.93182	2.72727
MAPE	9.02760	28.36502

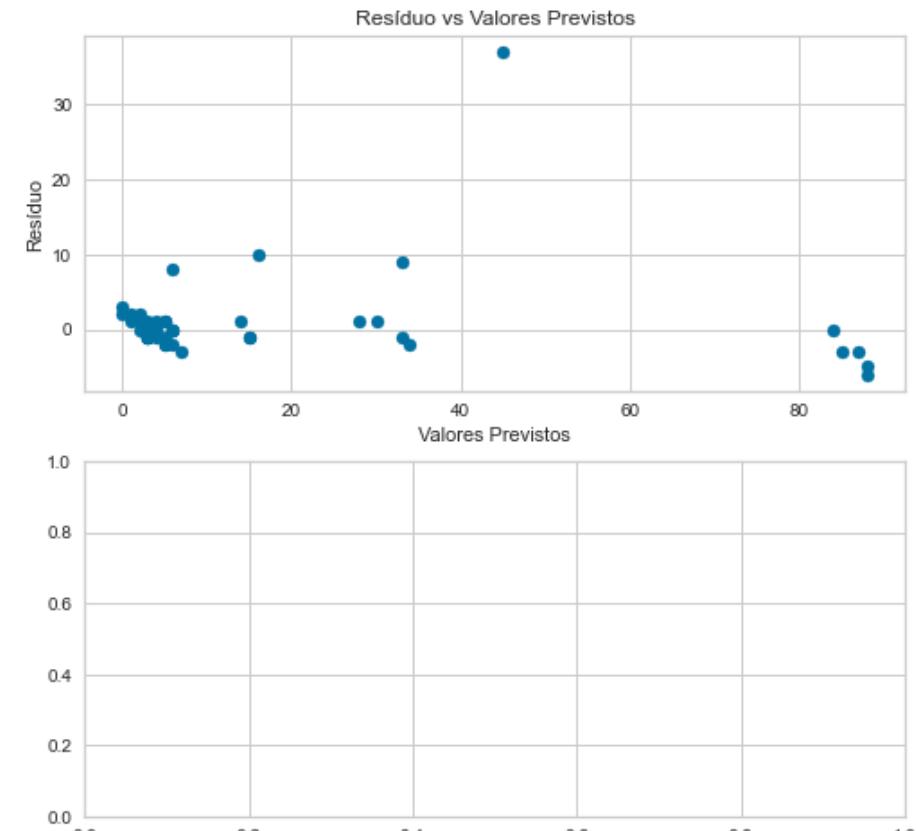
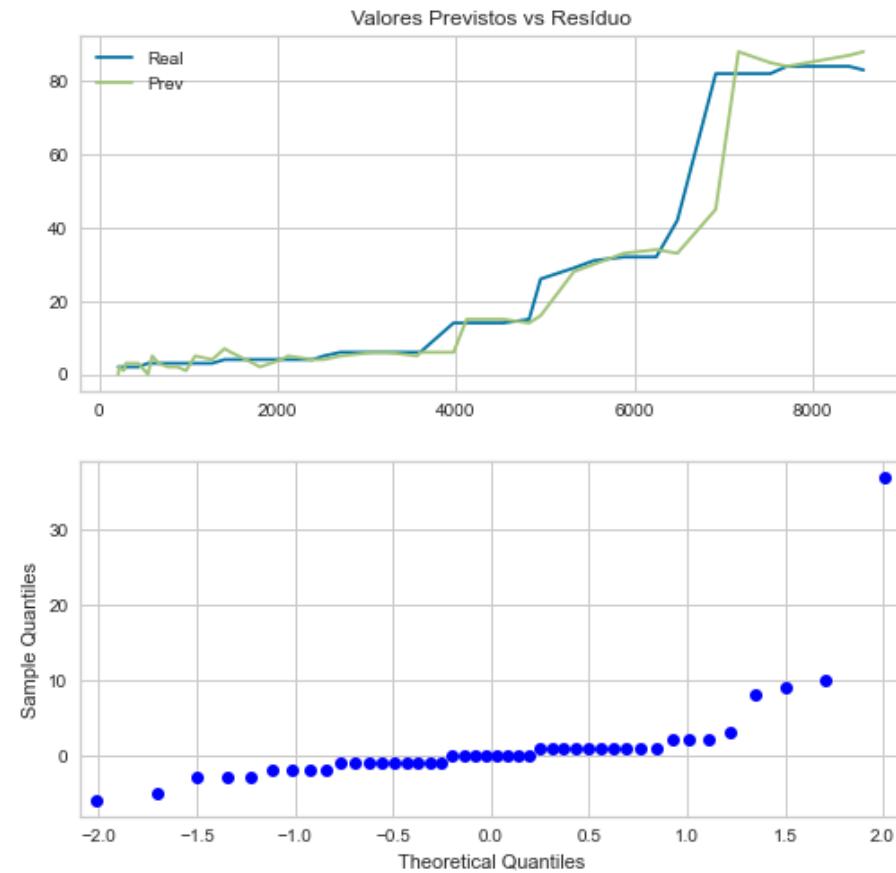
Com a **base teste**, a regressão de Theil-Sen em conjunto com o script de substituição de negativos pela lag de 1 dia teve **desempenho inferior à baseline**. O erro médio absoluto foi de 2.7 casos, contra 1.9 casos na baseline (que apenas repete o resultado do dia anterior). Quanto menor a quantidade de dados, pior é a previsão.



### 7.1.1. Método de Previsão em Laço

REGRESSÃO DE THEIL-SEN | CASOS ACUMULADOS

BASE TESTE – MUNICÍPIO DE SÃO SEBASTIÃO



## 7.1.2. Método de Previsão em Laço

REGRESSÃO DE THEIL-SEN | MORTES ACUMULADAS

191



PREVISÃO  
EM LAÇO

REGRESSÃO  
DE THEIL-  
SEN

MORTES  
ACUMULADAS



## 7.1.2. Método de Previsão em Laço

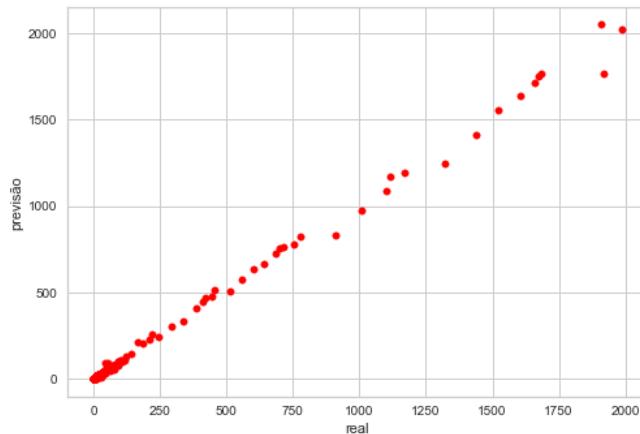
REGRESSÃO DE THEIL-SEN | MORTES ACUMULADAS

192

### BASE TESTE – MUNICÍPIOS DO ESTADO DE SP

#### Regressão de Theil-Sen + substituição de previsões negativas pela lag de 1 dia:

- Real | Previsto:
  - Média: 5.53411 | 5.59131
  - min: 0 | 0
  - 25%: 0 | 0
  - 50%: 0 | 0
  - 75%: 1 | 1
  - max: 1986 | 2049
- sem previsões negativas



Indicadores para base de teste	BASELINE	Theil-Sen Regressor + Baseline para Previsões Negativas (laço)
VIÉS	0.36453	-0.05720
MSE	21.03945	12.82113
RMSE	4.58688	3.58066
MAE	0.38204	0.37105
MAPE	nan	inf

Com a ABT com normalização robusta e a regressão de Theil-Sen, o laço funcionou durante todos os **128 dias**.

Com a **base teste**, a regressão de Theil-Sen em conjunto com o script de substituição de negativos pela lag de 1 dia teve **desempenho superior à baseline**. O erro médio absoluto foi de 0.37 mortes, contra 0.38 mortes na baseline (que apenas repete o resultado do dia anterior).

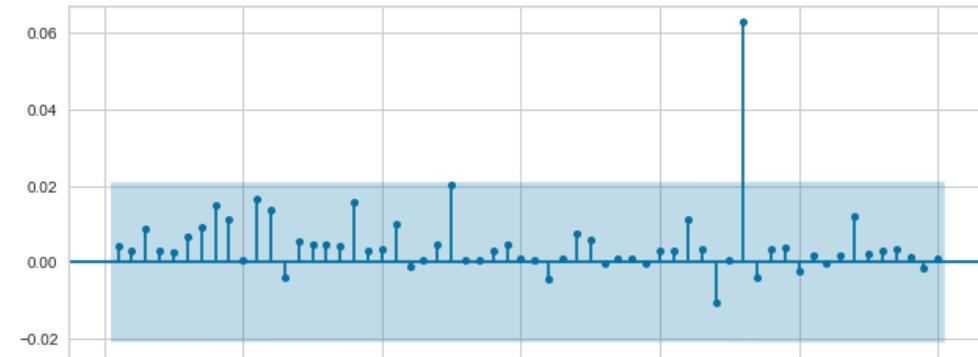
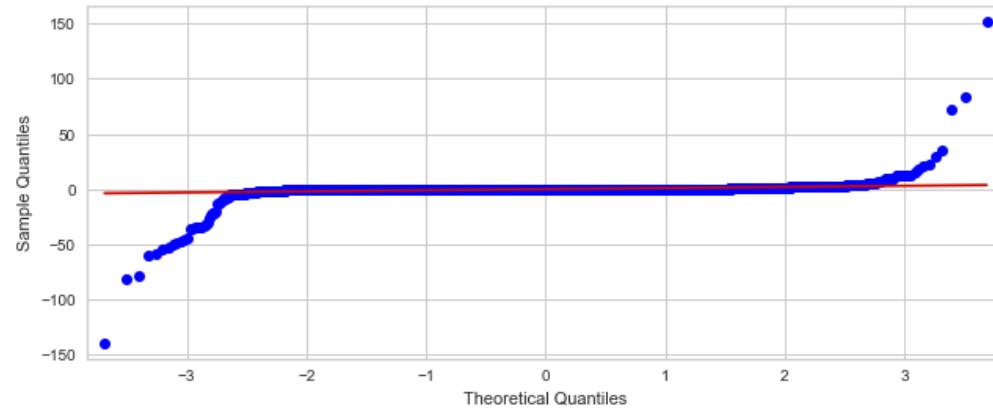
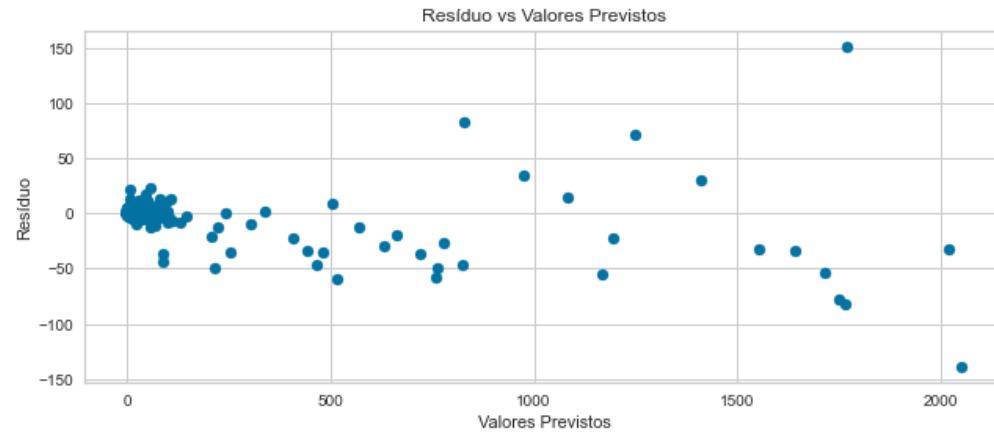


## 7.1.2. Método de Previsão em Laço

REGRESSÃO DE THEIL-SEN | MORTES ACUMULADAS

193

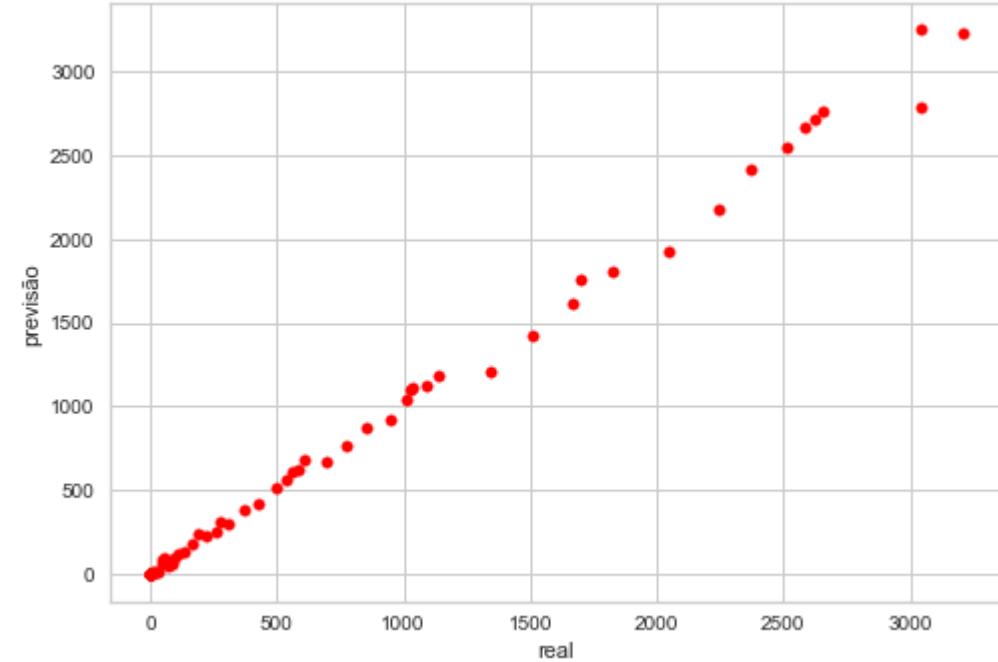
BASE TESTE – MUNICÍPIOS DO ESTADO DE SP



## 7.1.1. Método de Previsão em Laço

REGRESSÃO DE THEIL-SEN | CASOS ACUMULADOS

194



BASE TESTE – ESTADO DE SP

Indicadores para base de teste	BASELINE	Theil-Sen Regressor + Baseline para Previsões Negativas (laço)
VIÉS	45.15493	-7.08451
MSE	6808.33803	3244.29577
RMSE	82.51265	56.95872
MAE	45.15493	32.12676
MAPE	nan	15.46883

Com a **base teste**, a regressão de Theil-Sen em conjunto com o script de substituição de negativos pela lag de 1 dia teve **desempenho superior à baseline**. O erro médio absoluto foi de 32 mortes, contra 45 mortes na baseline (que apenas repete o resultado do dia anterior).

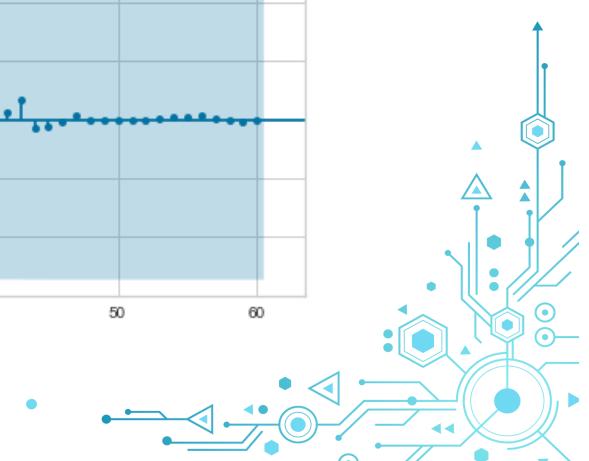
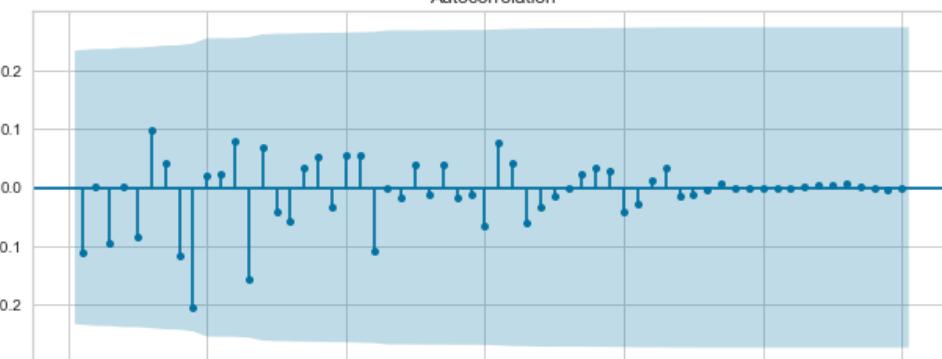
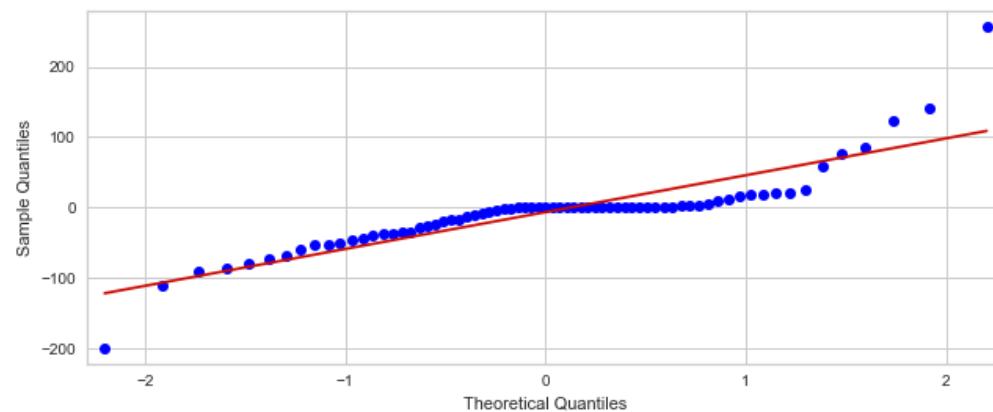
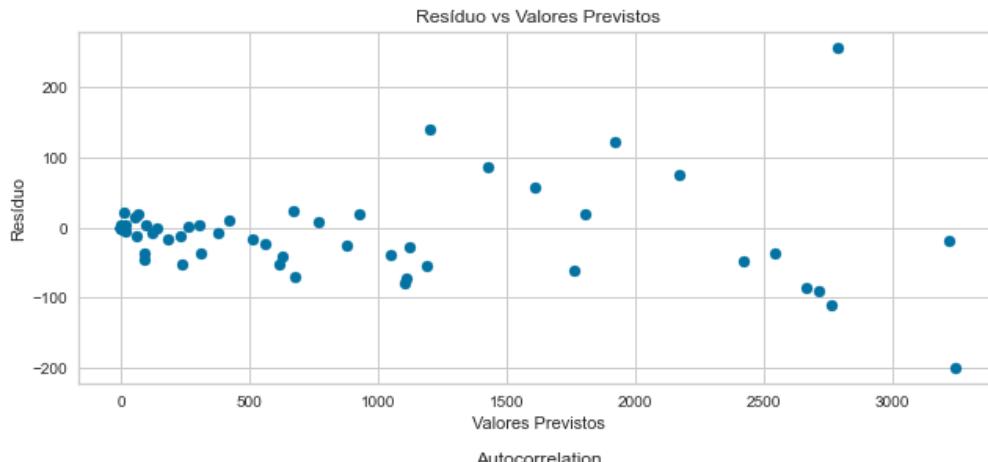
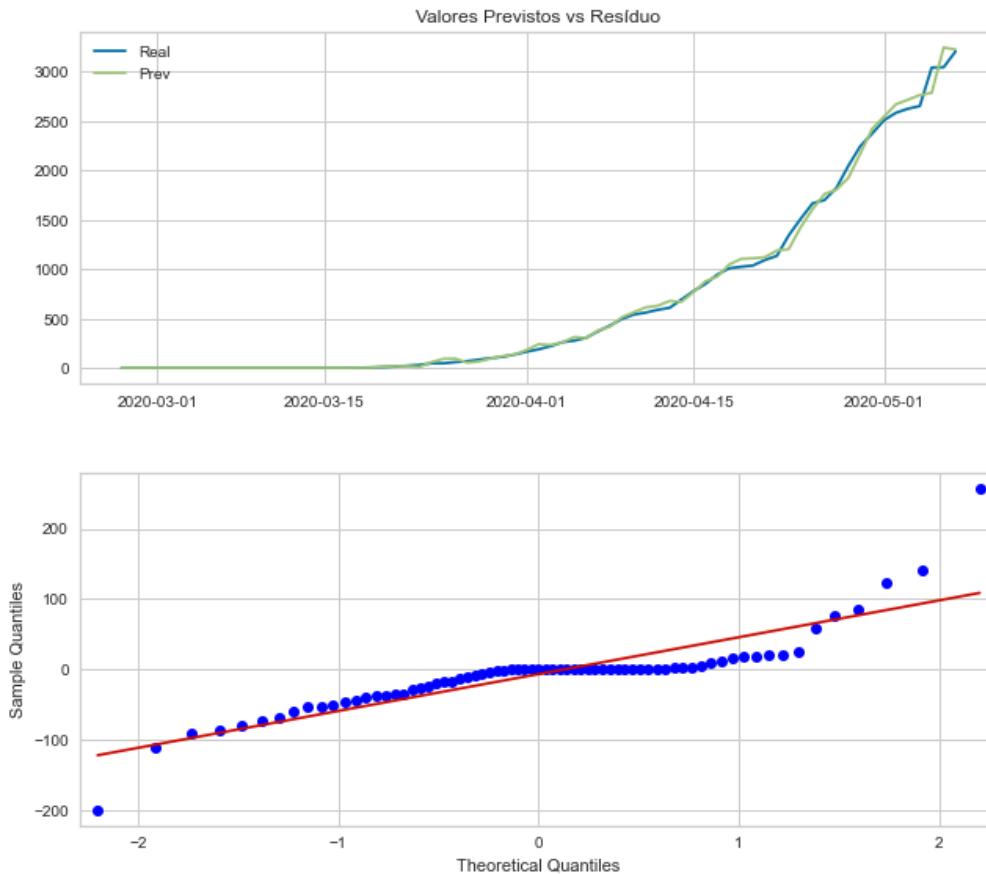


## 7.1.2. Método de Previsão em Laço

REGRESSÃO DE THEIL-SEN | MORTES ACUMULADAS

195

BASE TESTE – ESTADO DE SP

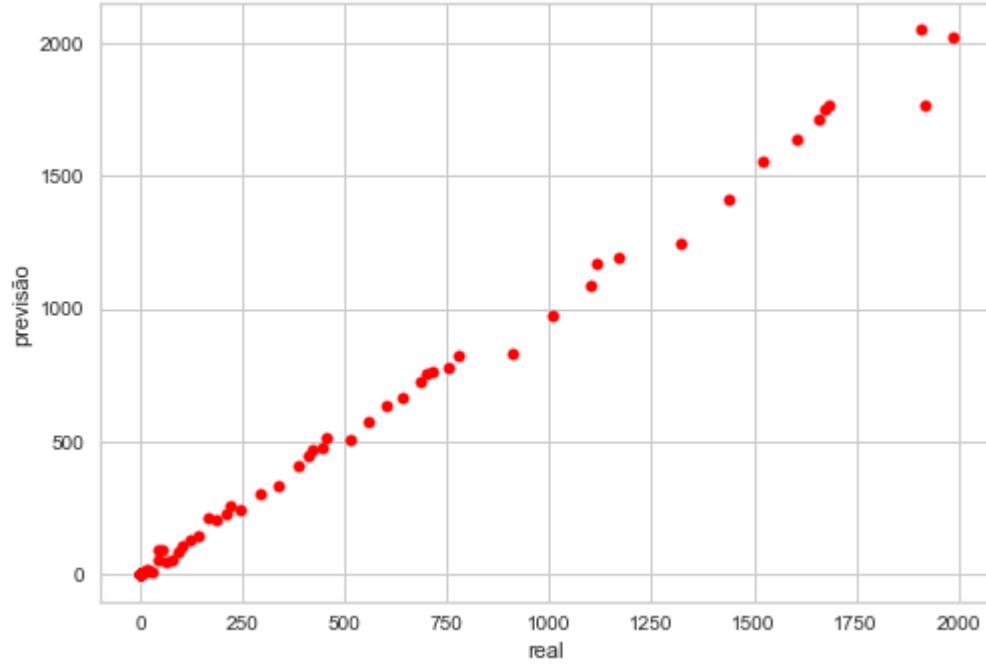


## 7.1.1. Método de Previsão em Laço

REGRESSÃO DE THEIL-SEN | CASOS ACUMULADOS

196

BASE TESTE – MUNICÍPIO DE SÃO PAULO



Indicadores para base de teste	BASELINE	Theil-Sen Regressor + Baseline para Previsões Negativas (laço)
VIÉS	27.97183	-10.94366
MSE	2527.66197	1526.01408
RMSE	50.27586	39.06423
MAE	28.22535	24.26761
MAPE	nan	16.39800

Com a **base teste**, a regressão de Theil-Sen em conjunto com o script de substituição de negativos pela lag de 1 dia teve **desempenho superior à baseline**. O erro médio absoluto foi de 24 mortes, contra 28 mortes na baseline (que apenas repete o resultado do dia anterior).

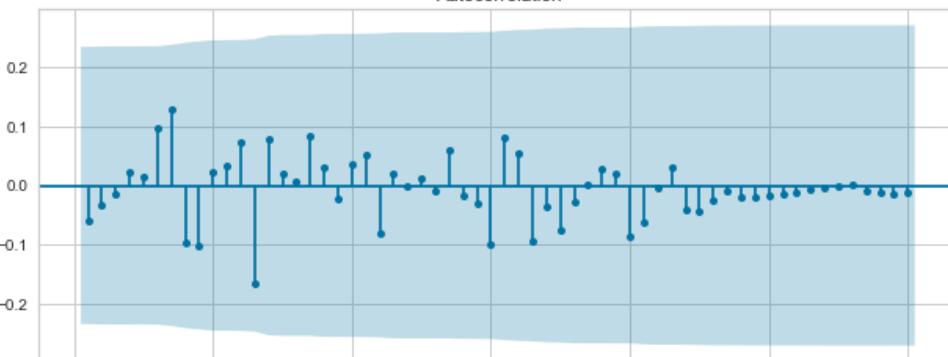
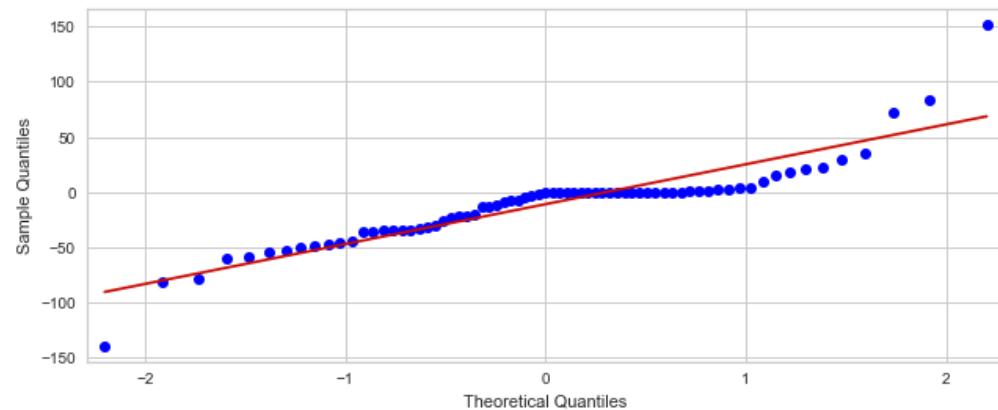
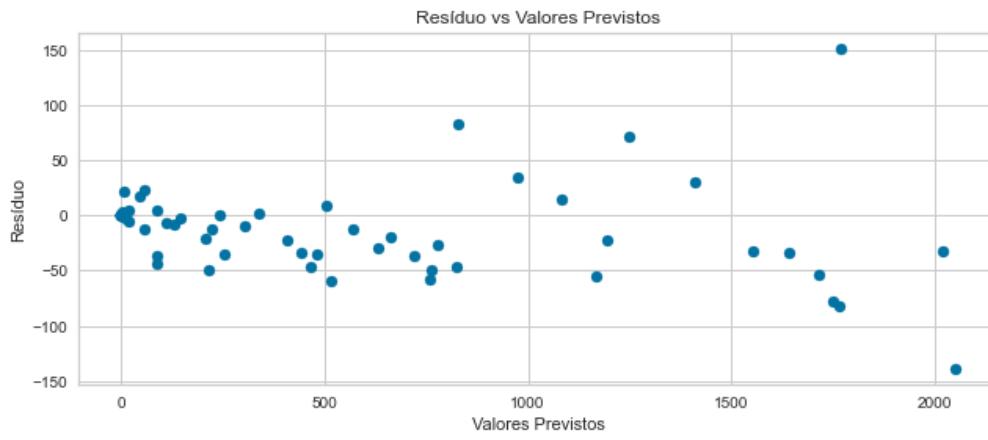
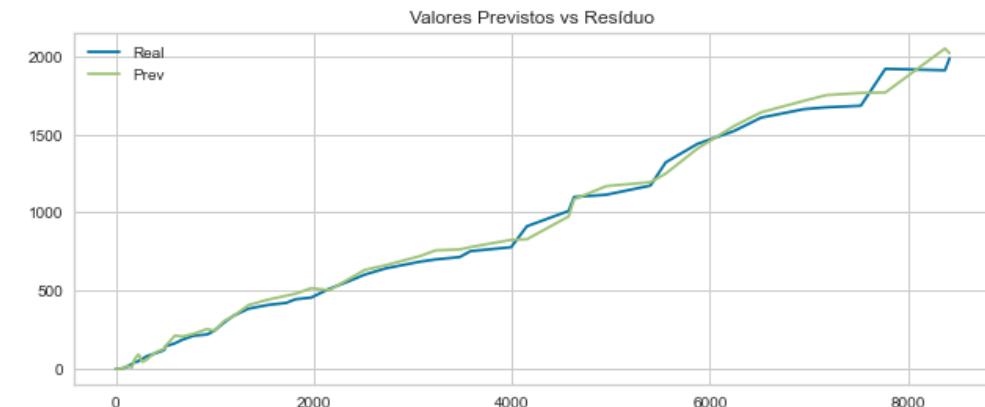


## 7.1.2. Método de Previsão em Laço

REGRESSÃO DE THEIL-SEN | MORTES ACUMULADAS

197

BASE TESTE – MUNICÍPIO DE SÃO PAULO

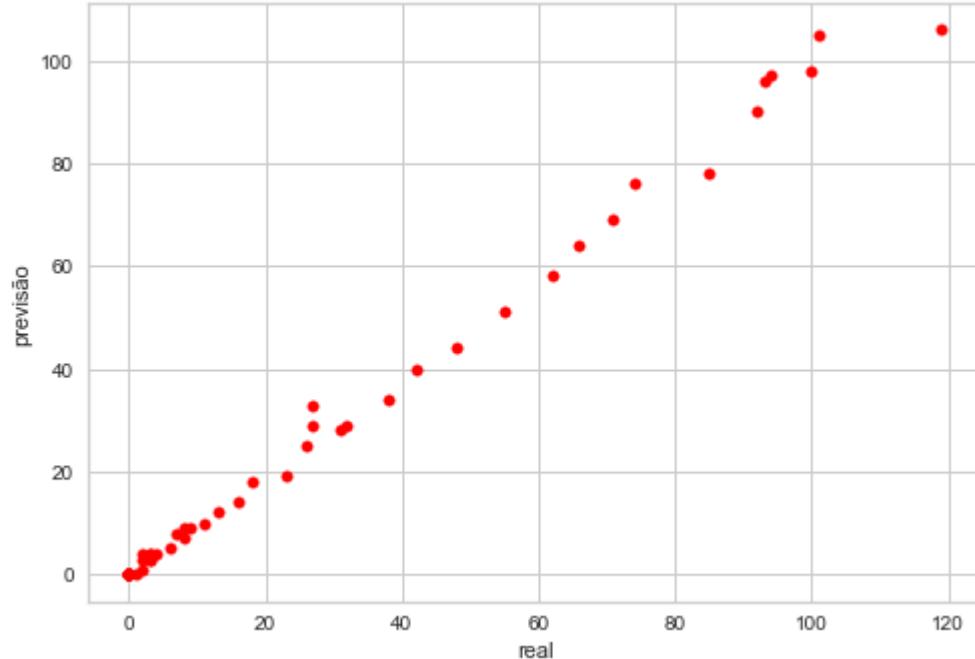


## 7.1.1. Método de Previsão em Laço

REGRESSÃO DE THEIL-SEN | CASOS ACUMULADOS

198

BASE TESTE – MUNICÍPIO DE OSASCO



Indicadores para base de teste	BASELINE	Theil-Sen Regressor + Baseline para Previsões Negativas (laço)
VIÉS	2.38000	0.76000
MSE	18.06000	8.68000
RMSE	4.24971	2.94618
MAE	2.58000	1.84000
MAPE	nan	15.42125

Com a **base teste**, a regressão de Theil-Sen em conjunto com o script de substituição de negativos pela lag de 1 dia teve **desempenho superior à baseline**. O erro médio absoluto foi de 1.84 mortes, contra 2.58 mortes na baseline (que apenas repete o resultado do dia anterior).

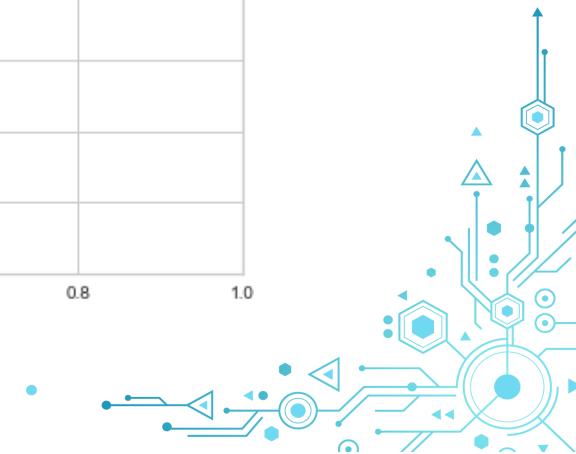
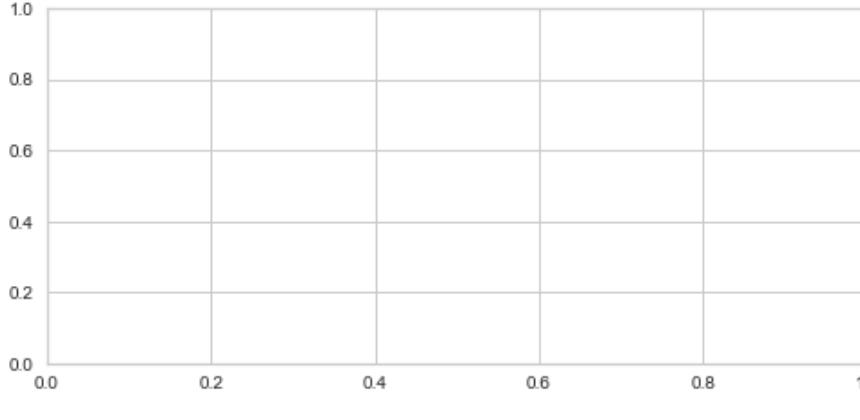
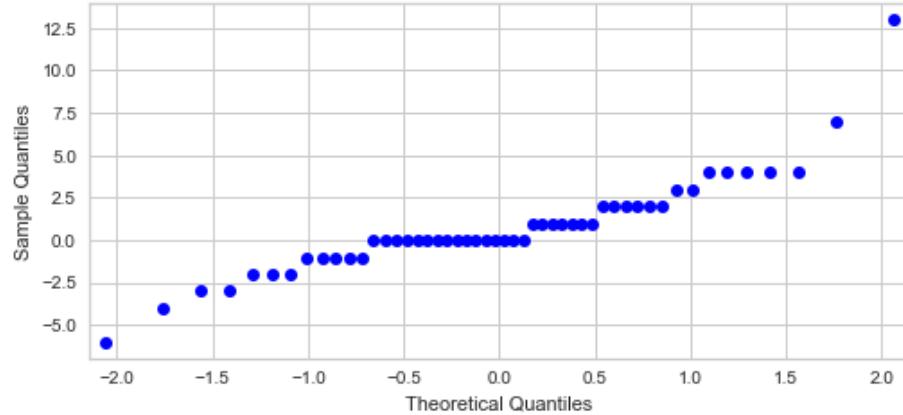
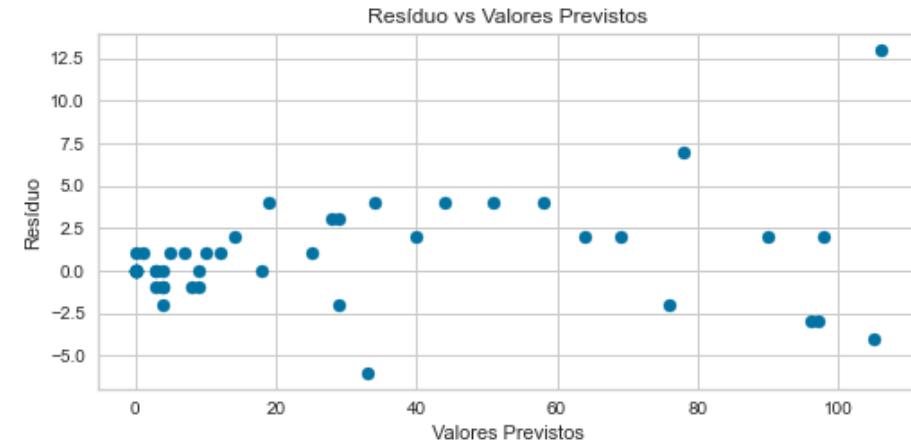
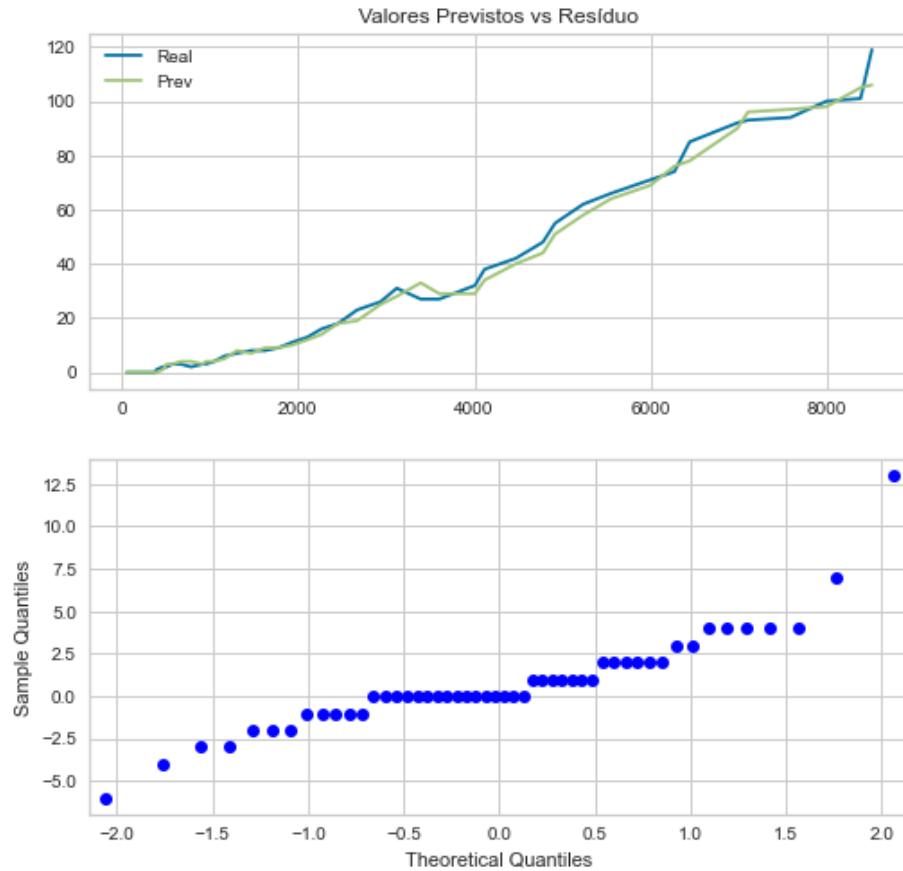


## 7.1.2. Método de Previsão em Laço

REGRESSÃO DE THEIL-SEN | MORTES ACUMULADAS

199

BASE TESTE – MUNICÍPIO DE OSASCO

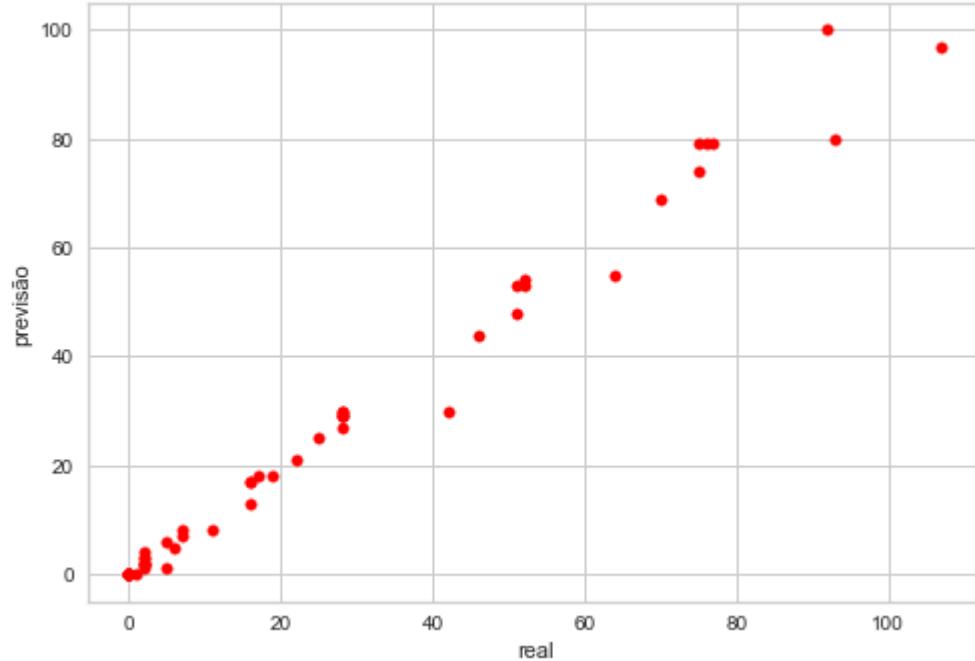


## 7.1.1. Método de Previsão em Laço

REGRESSÃO DE THEIL-SEN | CASOS ACUMULADOS

200

BASE TESTE – MUNICÍPIO DE GUARULHOS



Indicadores para base de teste	BASELINE	Theil-Sen Regressor + Baseline para Previsões Negativas (laço)
VIÉS	2.05769	0.55769
MSE	19.48077	12.94231
RMSE	4.41370	3.59754
MAE	2.09615	2.01923
MAPE	nan	16.11517

Com a **base teste**, a regressão de Theil-Sen em conjunto com o script de substituição de negativos pela lag de 1 dia teve **desempenho superior à baseline**. O erro médio absoluto foi de 2.01 mortes, contra 2.09 mortes na baseline (que apenas repete o resultado do dia anterior).

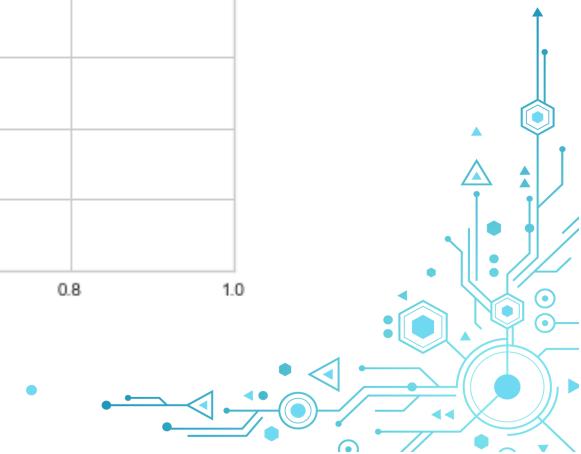
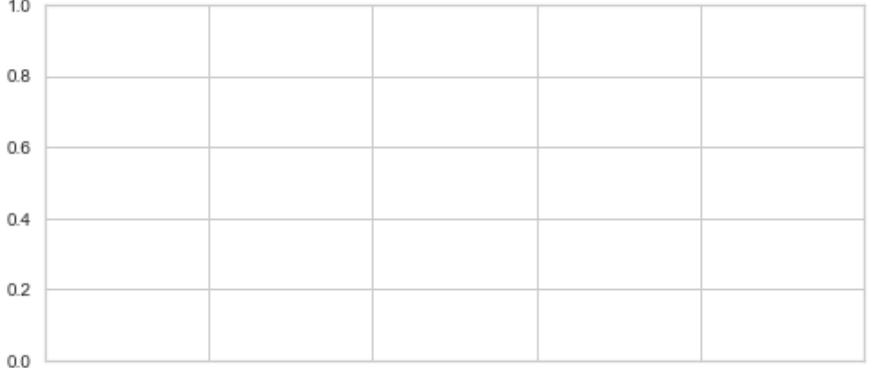
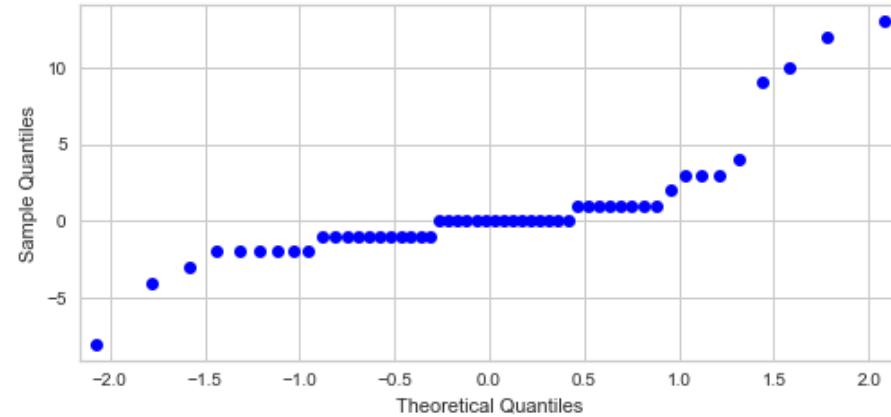
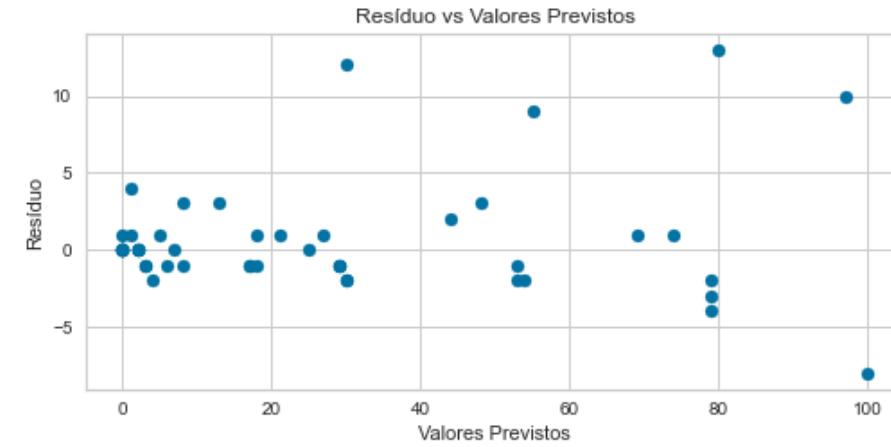
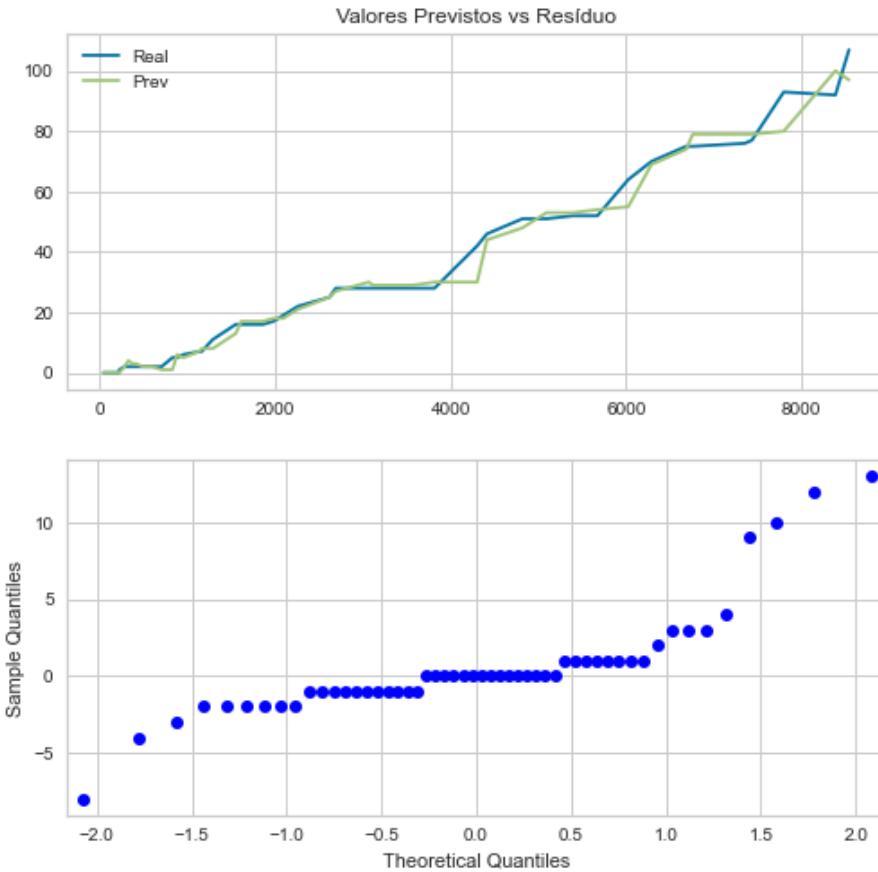


## 7.1.2. Método de Previsão em Laço

REGRESSÃO DE THEIL-SEN | MORTES ACUMULADAS

201

BASE TESTE – MUNICÍPIO DE GUARULHOS

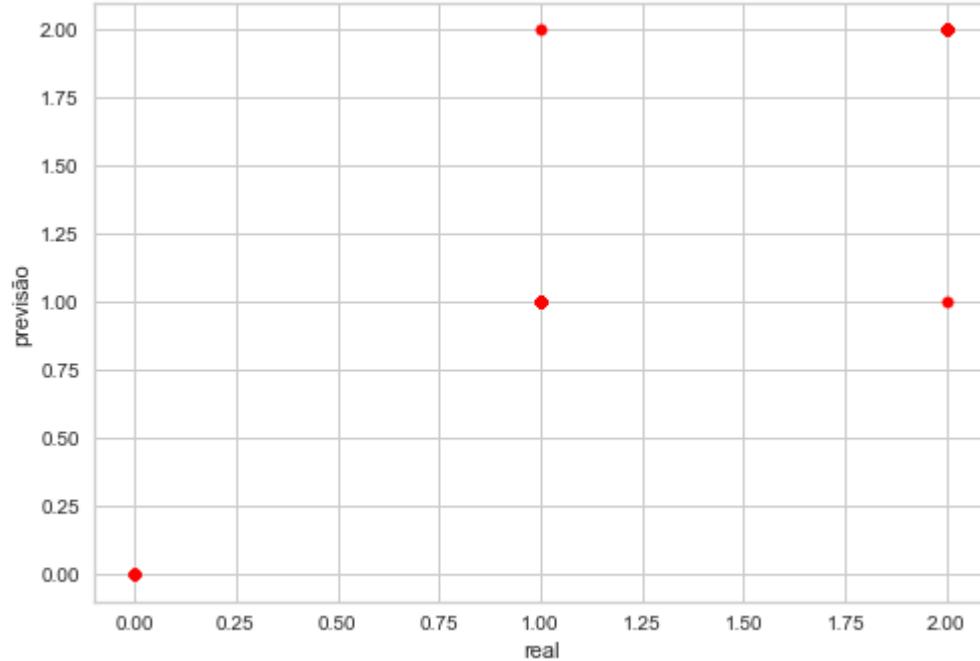


## 7.1.1. Método de Previsão em Laço

REGRESSÃO DE THEIL-SEN | CASOS ACUMULADOS

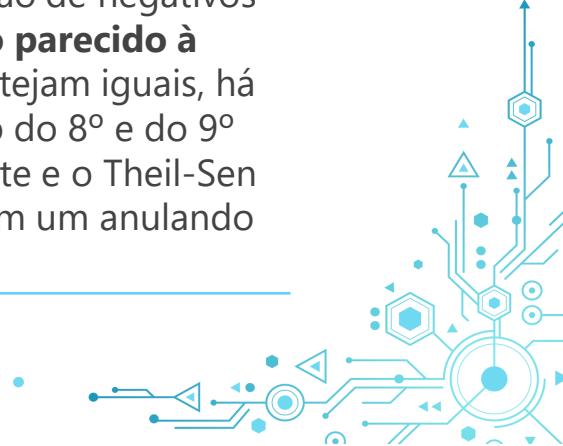
202

### BASE TESTE – MUNICÍPIO DE SÃO SEBASTIÃO



Indicadores para base de teste	BASELINE	Theil-Sen Regressor + Baseline para Previsões Negativas (laço)
VIÉS	0.04545	0.00000
MSE	0.04545	0.04545
RMSE	0.21320	0.21320
MAE	0.04545	0.04545
MAPE	nan	4.05405

Com a **base teste**, a regressão de Theil-Sen em conjunto com o script de substituição de negativos pela lag de 1 dia teve **desempenho parecido à baseline**. Embora os indicadores estejam iguais, há uma pequena diferença na previsão do 8º e do 9º dia: a baseline erra no 8º por 1 morte e o Theil-Sen erra no 9º também por 1 morte, com um anulando o outro.

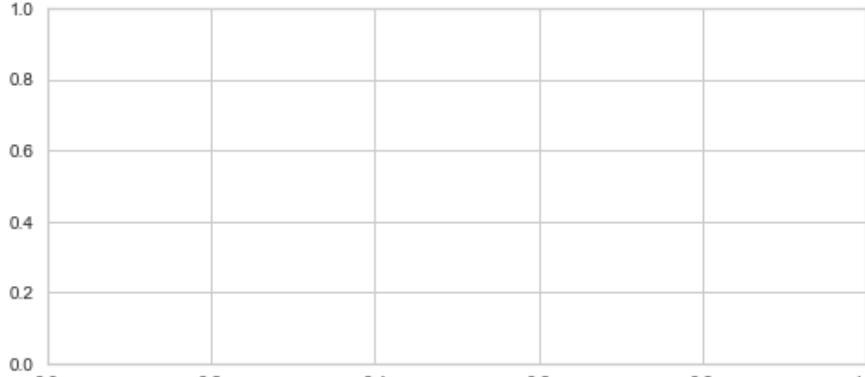
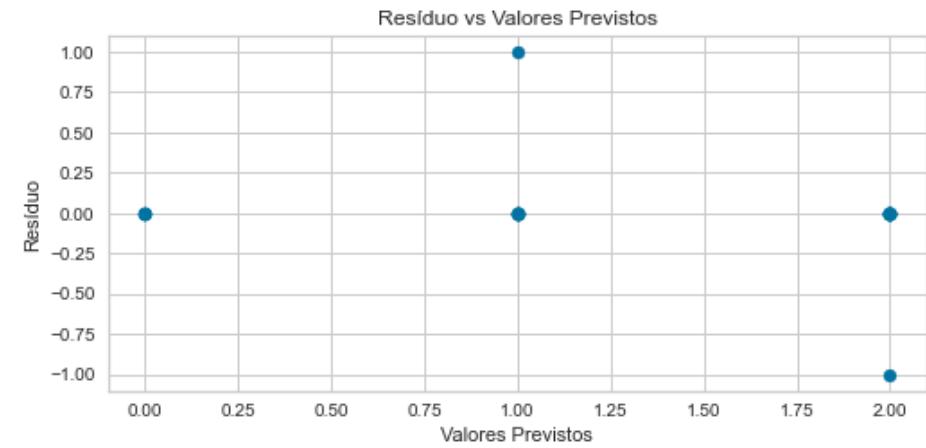
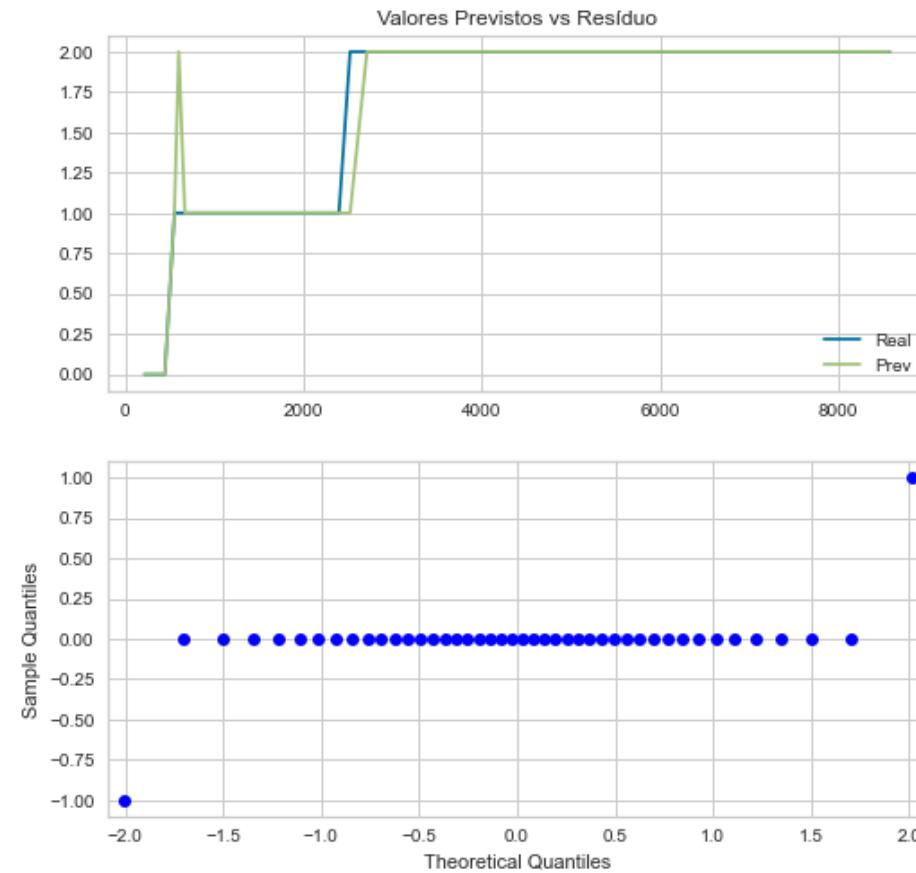


## 7.1.2. Método de Previsão em Laço

REGRESSÃO DE THEIL-SEN | MORTES ACUMULADAS

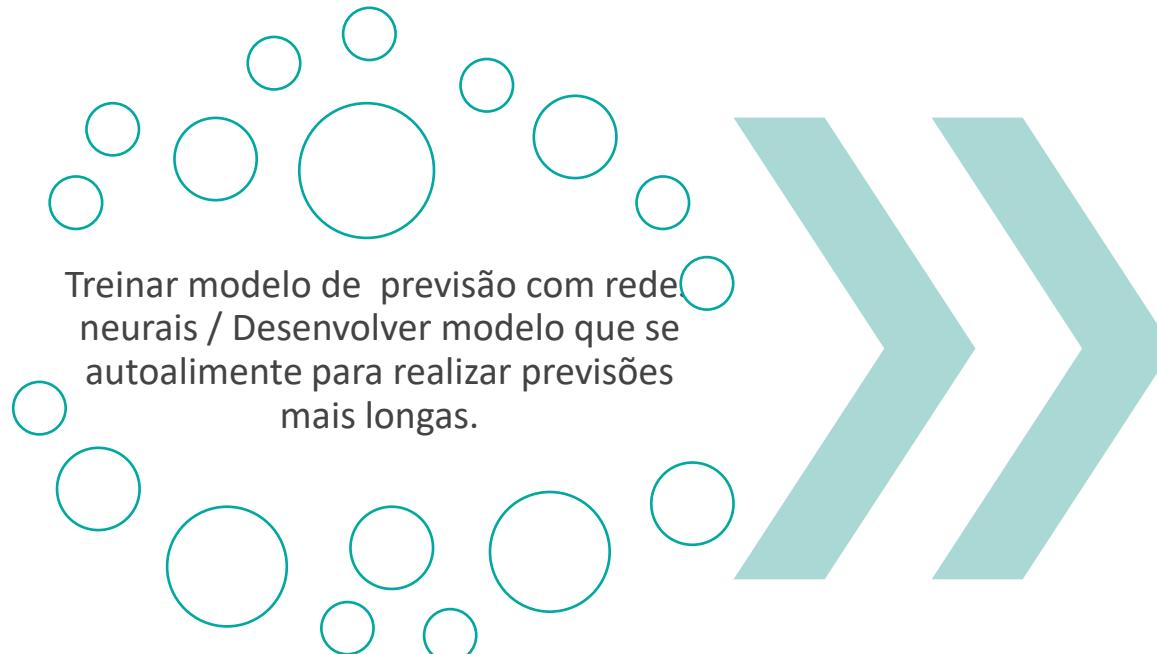
203

BASE TESTE – MUNICÍPIO DE SÃO SEBASTIÃO



## 8. Sugestão de Trabalhos Futuros

204



Monitorar desempenho do algoritmo com novos dados / Aplicar novamente o Pycaret quando o desempenho do algoritmo com novos dados igualar à baseline e definir novo algoritmo.



# LABDATA FIA – Laboratório de Análise de Dados



Unidade Pinheiros



Unidade Paulista



## - Análises Complementares -



# Detalhes das análises

'COD7D', 'CIDADE', 'MUNUF' | TOP 20

207

cod7d	munuf	qtd.	%
3550308	São Paulo-SP	72	0.818647
3547304	Santana de Parnaíba-SP	61	0.693576
3515707	Ferraz de Vasconcelos-SP	57	0.648096
3510609	Carapicuíba-SP	55	0.625355
3548708	São Bernardo do Campo-SP	54	0.613985
3547809	Santo André-SP	53	0.602615
3548807	São Caetano do Sul-SP	53	0.602615
3518800	Guarulhos-SP	52	0.591245
3529401	Mauá-SP	52	0.591245
3513009	Cotia-SP	51	0.579875
3509502	Campinas-SP	51	0.579875
3524709	Jaguariúna-SP	51	0.579875
3549904	São José dos Campos-SP	51	0.579875
3505708	Barueri-SP	51	0.579875
3549805	São José do Rio Preto-SP	51	0.579875
3552502	Suzano-SP	50	0.568505
3534401	Osasco-SP	50	0.568505
3554102	Taubaté-SP	50	0.568505
3530607	Mogi das Cruzes-SP	49	0.557135
3519071	Hortolândia-SP	49	0.557135



# Detalhes das análises

'COD7D', 'CIDADE', 'MUNUF' | BOTTOM 20

208

<b>cod7d</b>	<b>munuf</b>	<b>qtd.</b>	<b>%</b>
3546306	Santa Cruz das Palmeiras-SP	2	0.02274
3553609	Tapiratiba-SP	2	0.02274
3551207	Sarutaiá-SP	2	0.02274
3507753	Brejo Alegre-SP	2	0.02274
3552700	Tabatinga-SP	2	0.02274
3514106	Dois Córregos-SP	2	0.02274
3519105	Iacanga-SP	2	0.02274
3517505	Guapiaçu-SP	2	0.02274
3540507	Porangaba-SP	1	0.01137
3509957	Canas-SP	1	0.01137
3531407	Monte Aprazível-SP	1	0.01137
3553500	Tapiraí-SP	1	0.01137
3546256	Santa Cruz da Esperança-SP	1	0.01137
3516804	Gastão Vidigal-SP	1	0.01137
3501806	Américo de Campos-SP	1	0.01137
3555703	União Paulista-SP	1	0.01137
3521309	Ipuã-SP	1	0.01137
3518503	Guareí-SP	1	0.01137
3504909	Bananal-SP	1	0.01137
3550001	São Luiz do Paraitinga-SP	1	0.01137



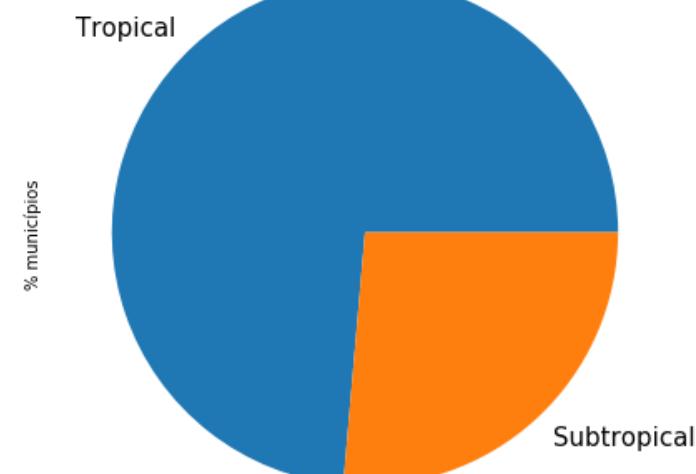
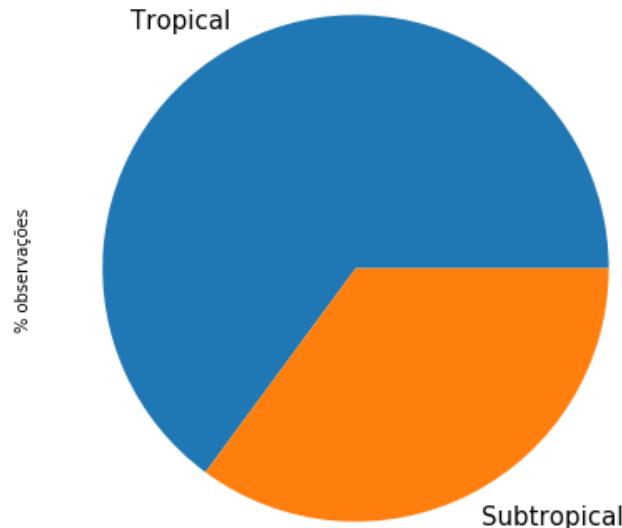
# 4. Análise Exploratória de Dados

RAIO-X DA BASE | ANÁLISE UNIVARIADA

209

## 'lat'

- Para latitude, não faz sentido calcular a frequência de observações pura e simples. Ao invés disso, será criada uma nova coluna para apontar a zona geográfica de acordo com a latitude (tropical, subtropical, temperada, polar). Mais a frente, nas análises multivariadas, latitude e longitude serão utilizadas para plotar no mapa de São Paulo os casos e as mortes confirmados de COVID-19.
- Quase 65% das observações estão na zona tropical, embora quase 3/4 dos municípios do dataset estejam nessa região, o que pode ser explicado pela epidemia de COVID-19 no Brasil ter se iniciado na região subtropical (SP) e portanto estar mais "madura" nessa região. Hipótese: com o tempo, a tendência é que a distribuição de observações nas zonas se aproxime mais da distribuição dos municípios nas zonas. Se isso não se comprovar, pode ser um indicativo de que o clima afeta o contágio do novo coronavírus.

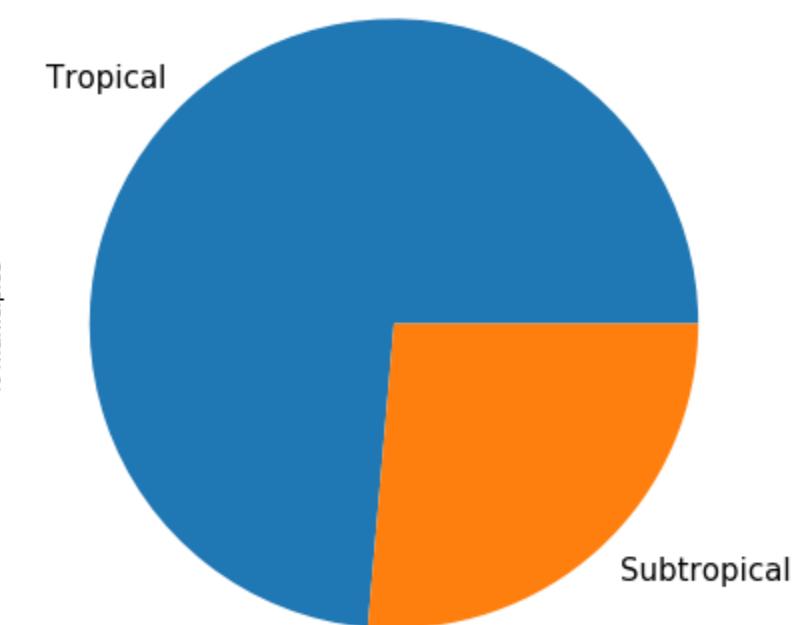
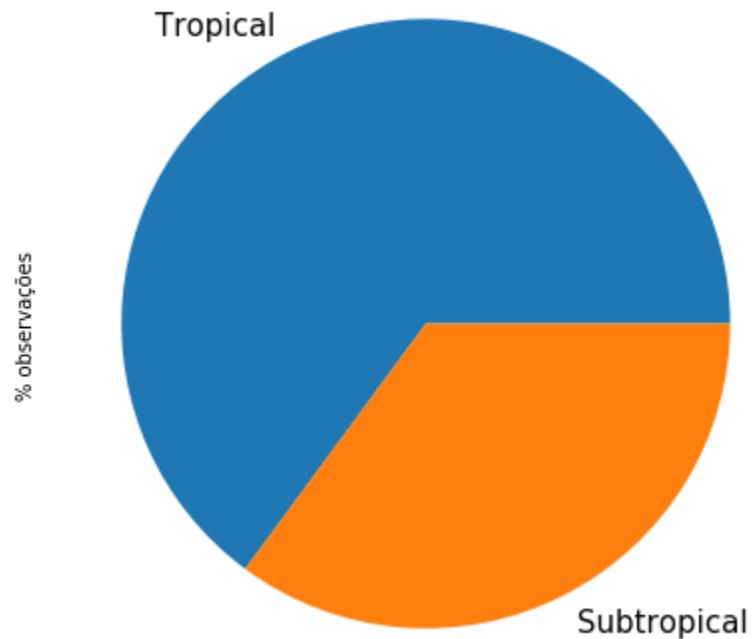


# Detalhes das análises

'LAT' | DISTRIBUIÇÃO DE OBSERVAÇÕES E DE MUNICÍPIOS POR FAIXA DE LATITUDE

210

zona latitudinal	qtd. observações	% observações	qtd. municípios	% municípios
Tropical	5707	64.889142	282	73.629243
Subtropical	3088	35.110858	101	26.370757



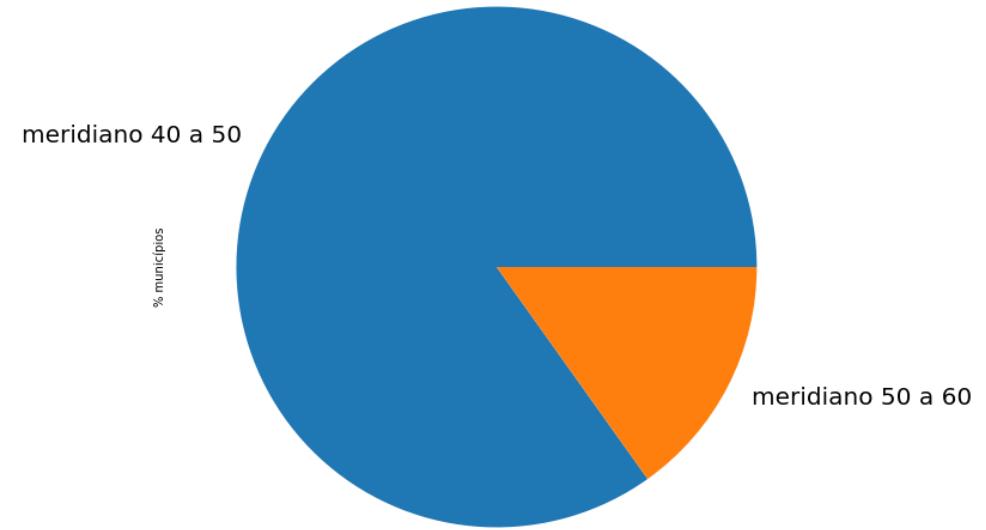
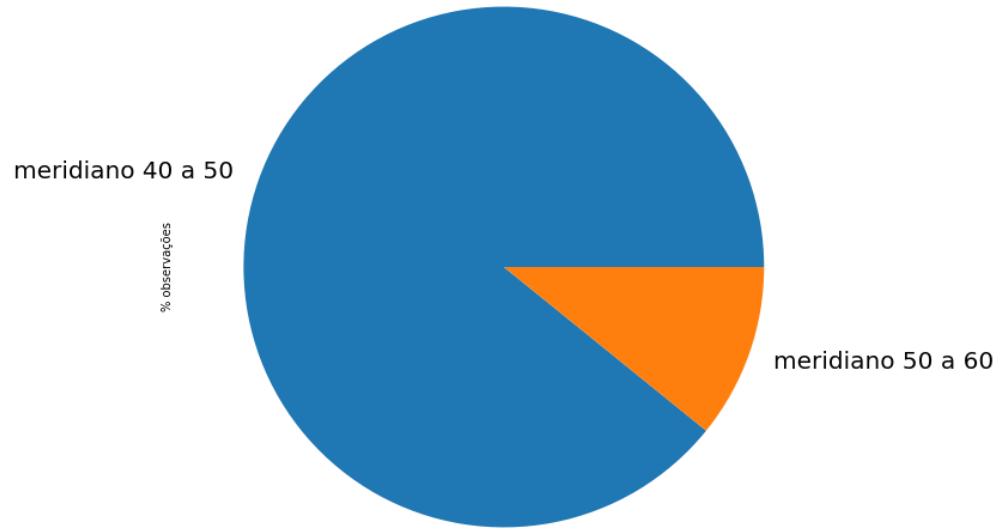
# 4. Análise Exploratória de Dados

RAIO-X DA BASE | ANÁLISE UNIVARIADA

211

## 'lon'

- Para longitude, não faz sentido calcular a frequência de observações pura e simples. Ao invés disso, será criada uma nova coluna para apontar em qual faixa de meridianos o município se encontra, contando de 10 em 10, para verificar a trajetória do vírus em direção ao interior do estado de São Paulo com o número de observações. Mais a frente, nas análises multivariadas, latitude e longitude serão utilizadas para plotar no mapa de São Paulo os casos e as mortes confirmados de COVID-19.
- Ao contrário da análise por latitude, a análise por longitude não traz grande diferença na distribuição de observações por faixa longitudinal e a distribuição de municípios por faixa longitudinal. Mesmo assim, ainda nota-se ligeira vantagem no percentual de observações em relação ao percentual de municípios na faixa longitudinal 40, o que parece fazer sentido, já que o contágio começou na faixa de 40 a 50 e seguiu para o interior.

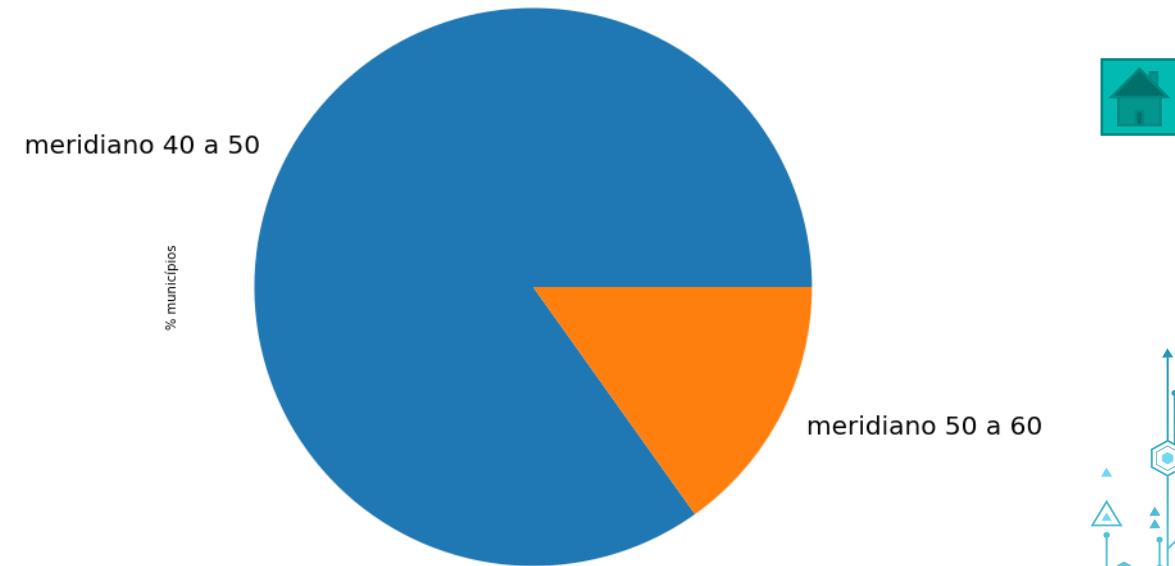
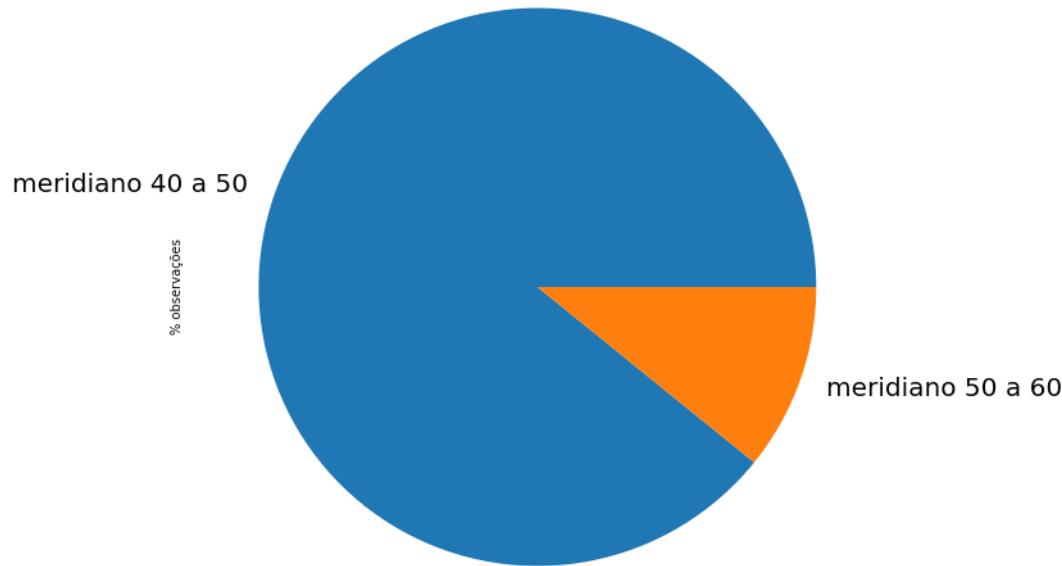


# Detalhes das análises

'LON' | DISTRIBUIÇÃO DE OBSERVAÇÕES E DE MUNICÍPIOS POR FAIXA DE LONGITUDE

212

<b>zona longitudinal</b>	<b>qtd. observações</b>	<b>% observações</b>	<b>qtd. municípios</b>	<b>% municípios</b>
meridiano 40 a 50	7843	89.175668	325	84.856397
meridiano 50 a 60	952	10.824332	58	15.143603



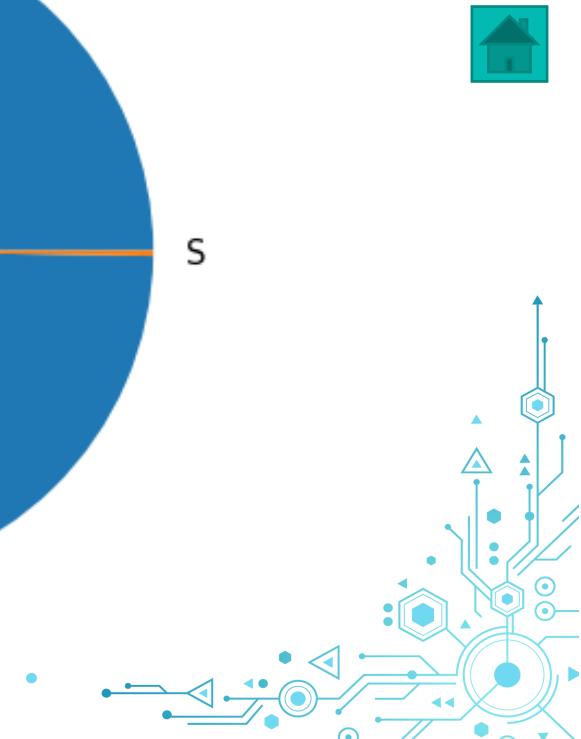
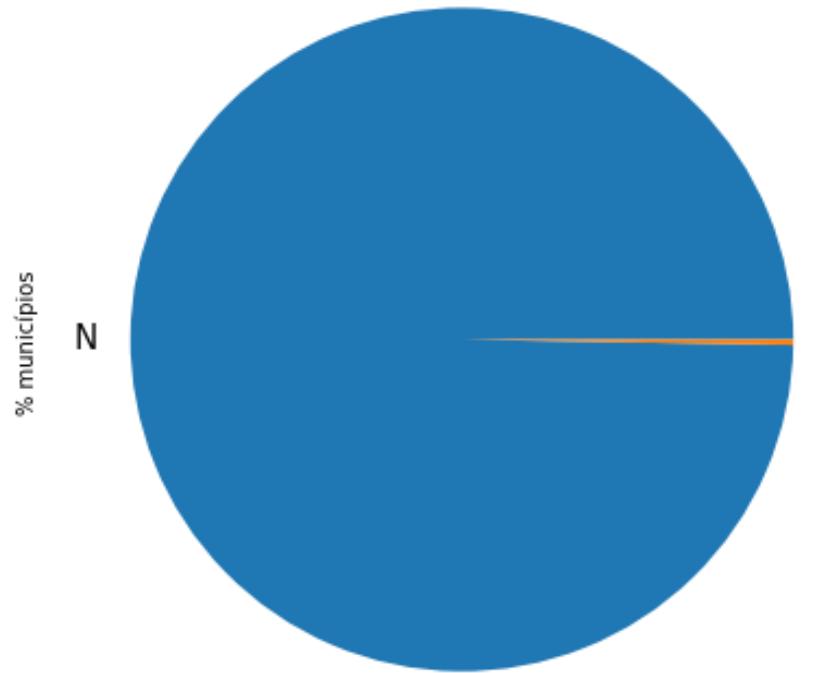
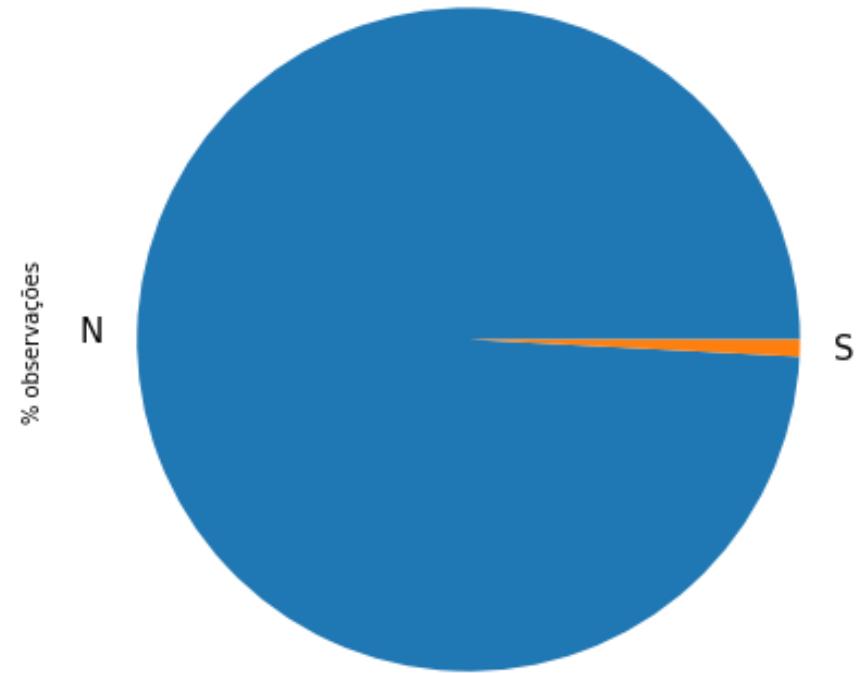
## 4. Análise Exploratória de Dados

RAIO-X DA BASE | ANÁLISE UNIVARIADA

213

'capital'

- A maior concentração proporcional de observações na capital demonstra que até a data de corte, a COVID-19 tinha mais presença na capital do que no interior.

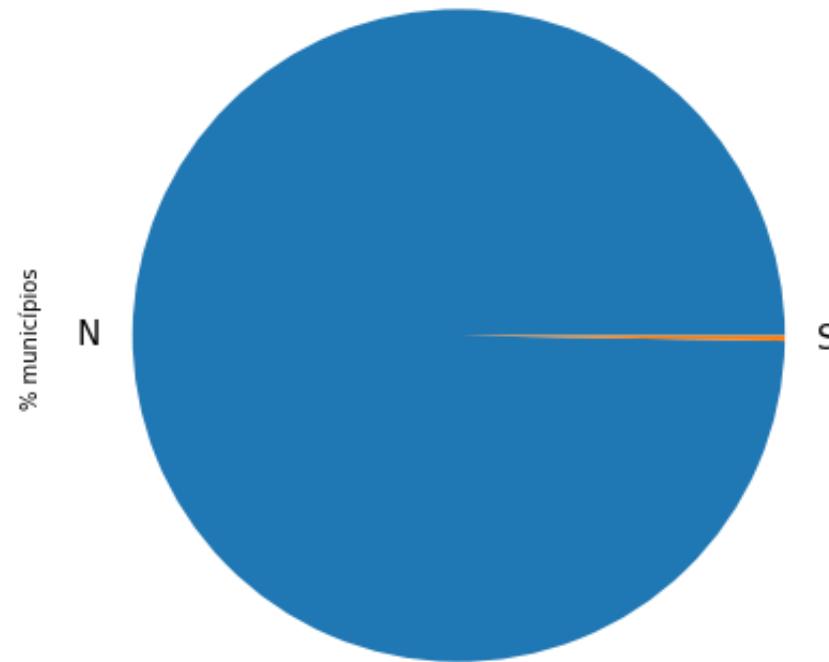
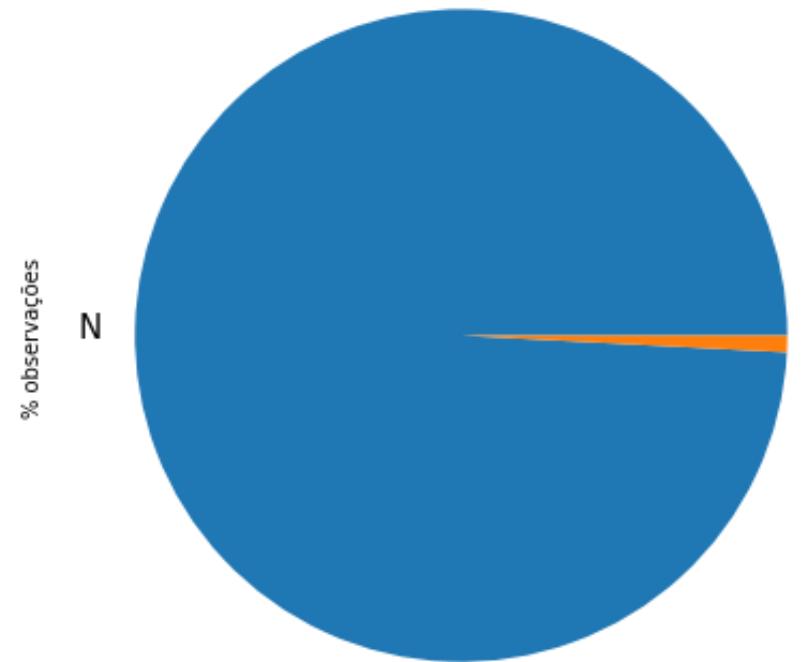


# Detalhes das análises

'CAPITAL' | DISTRIBUIÇÃO DE OBSERVAÇÕES E DE MUNICÍPIOS POR TIPO CAPITAL (SIM OU NÃO)

214

capital	qtd. observações	% observações	qtd. municípios	% municípios
N	8723	99.181353	382	99.738903
S	72	0.818647	1	0.261097



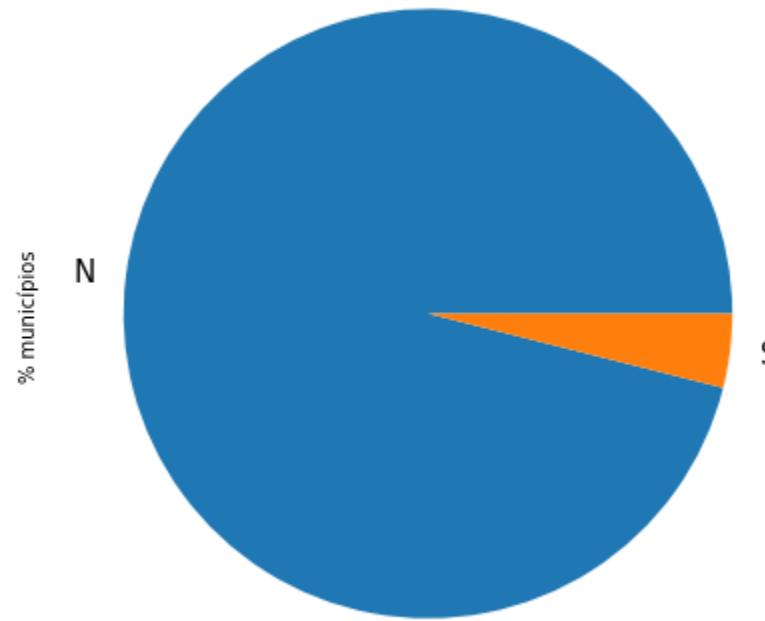
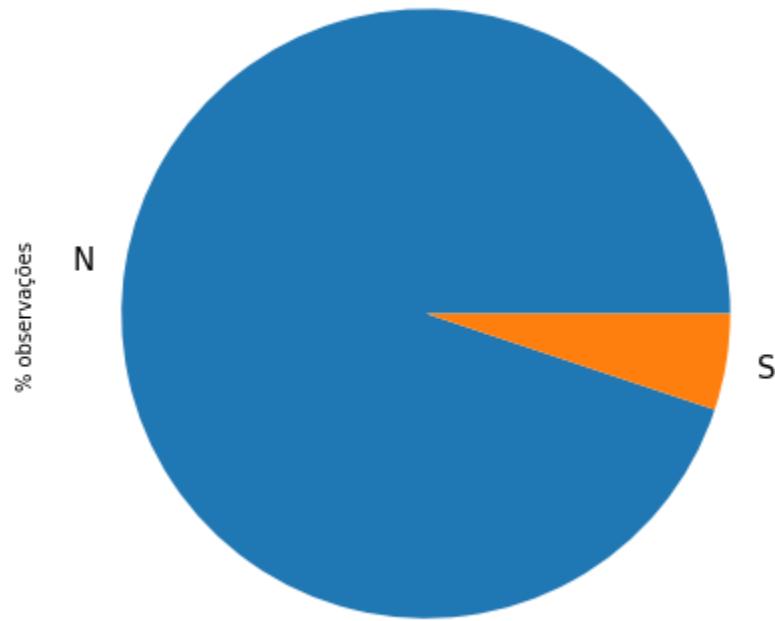
# 4. Análise Exploratória de Dados

RAIO-X DA BASE | ANÁLISE UNIVARIADA

215

## 'litoral'

- A maior concentração de observações no litoral, comparativamente ao percentual de municípios que estão na faixa litorânea, pode ser um indicativo de que o coronavírus teve taxa de transmissão mais alta no litoral, devido a condições socioeconômicas. Vale investigar isso nas análises multivariáveis.

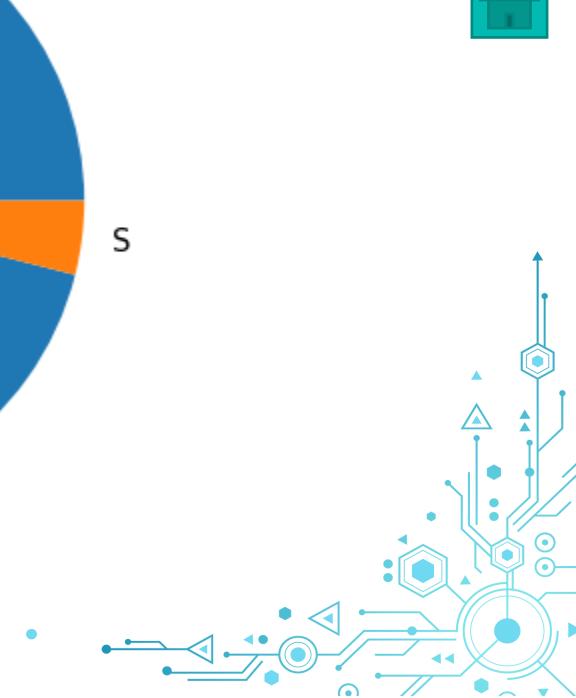
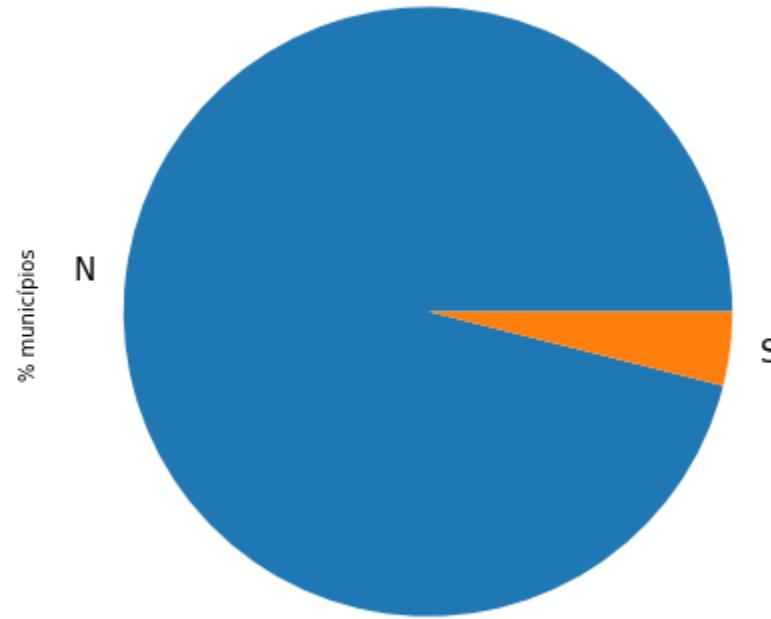
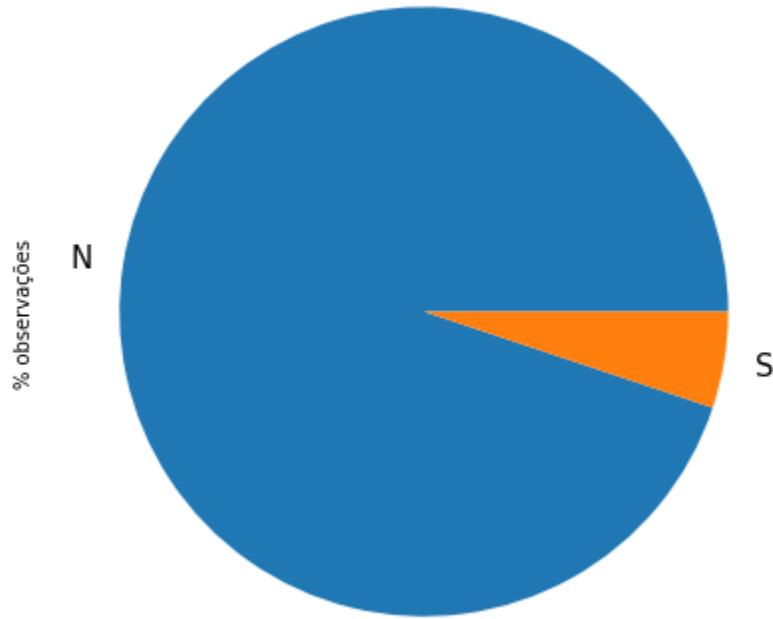


# Detalhes das análises

'LITORAL' | DISTRIBUIÇÃO DE OBSERVAÇÕES E DE MUNICÍPIOS POR TIPO LITORAL (SIM OU NÃO)

216

<b>litoral</b>	<b>qtd. observações</b>	<b>% observações</b>	<b>qtd. municípios</b>	<b>% municípios</b>
N	8346	94.894827	368	96.083551
S	449	5.105173	15	3.916449



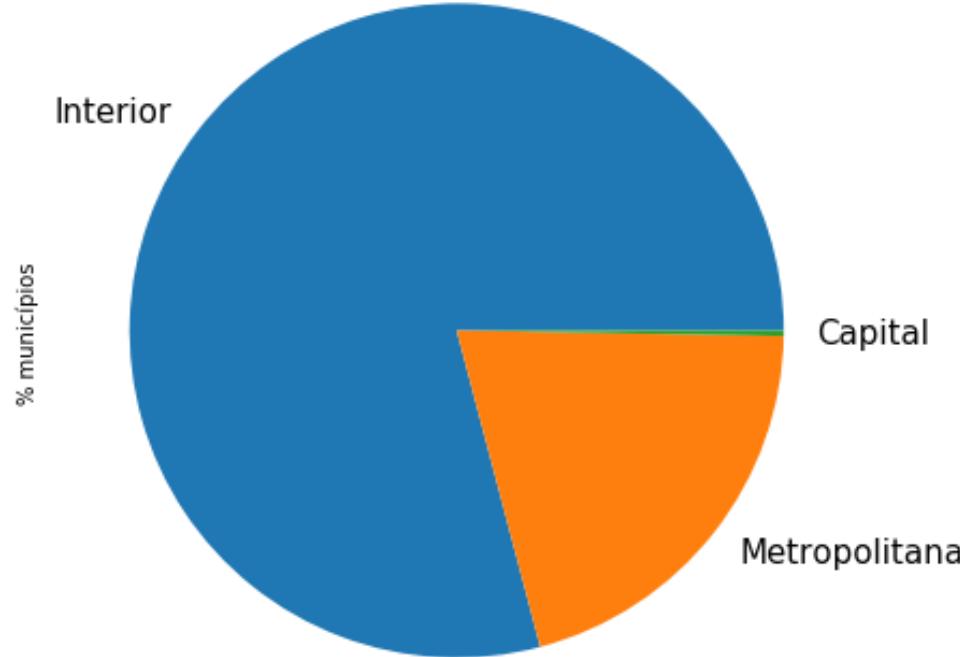
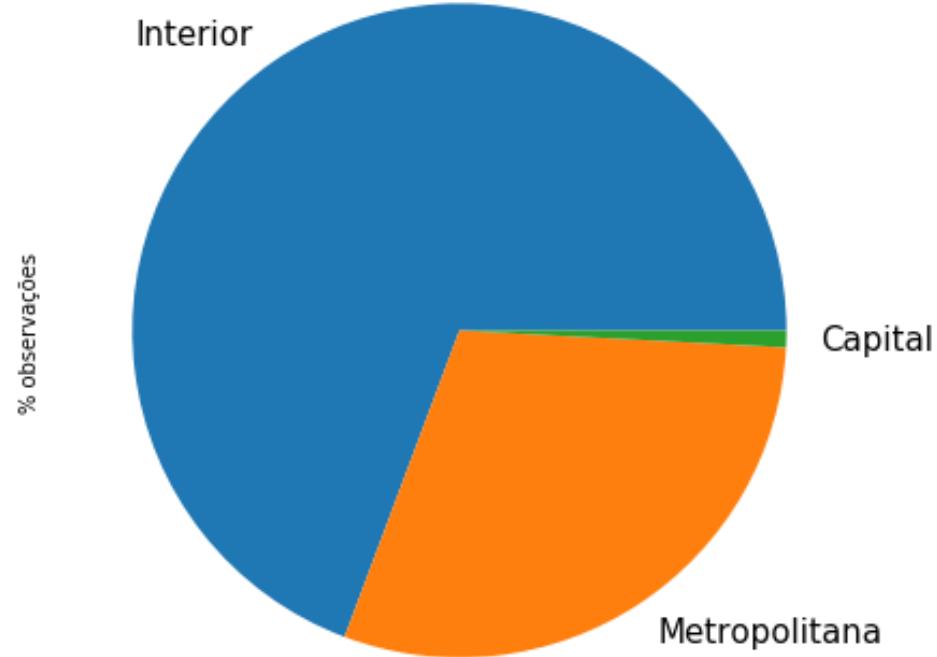
# 4. Análise Exploratória de Dados

RAIO-X DA BASE | ANÁLISE UNIVARIADA

217

'papel'

- Aqui se vê mais claramente que a pandemia ainda não havia avançado suficientemente no interior, como se pode notar pela maior distribuição relativa de observações na capital e nas regiões metropolitanas.

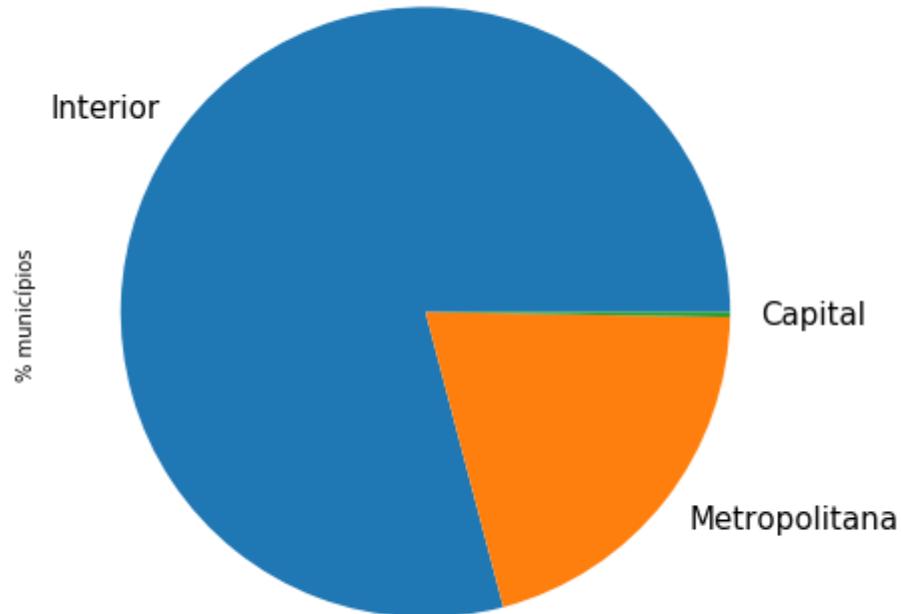
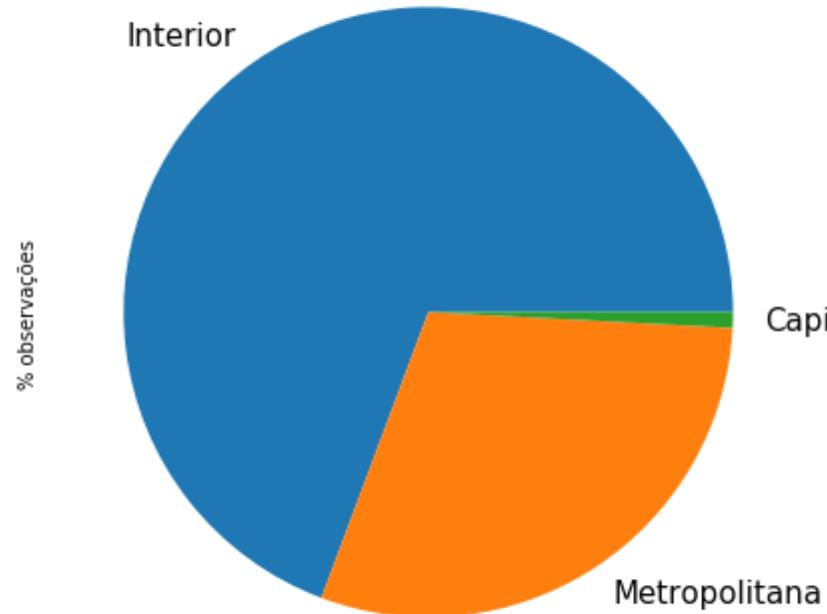


# Detalhes das análises

'PAPEL' | DISTRIBUIÇÃO DE OBSERVAÇÕES E DE MUNICÍPIOS POR PAPEL

218

papel	qtd. observações	% observações	qtd. municípios	% municípios
Interior	6094	69.289369	303	79.112272
Metropolitana	2629	29.891984	79	20.626632
Capital	72	0.818647	1	0.261097



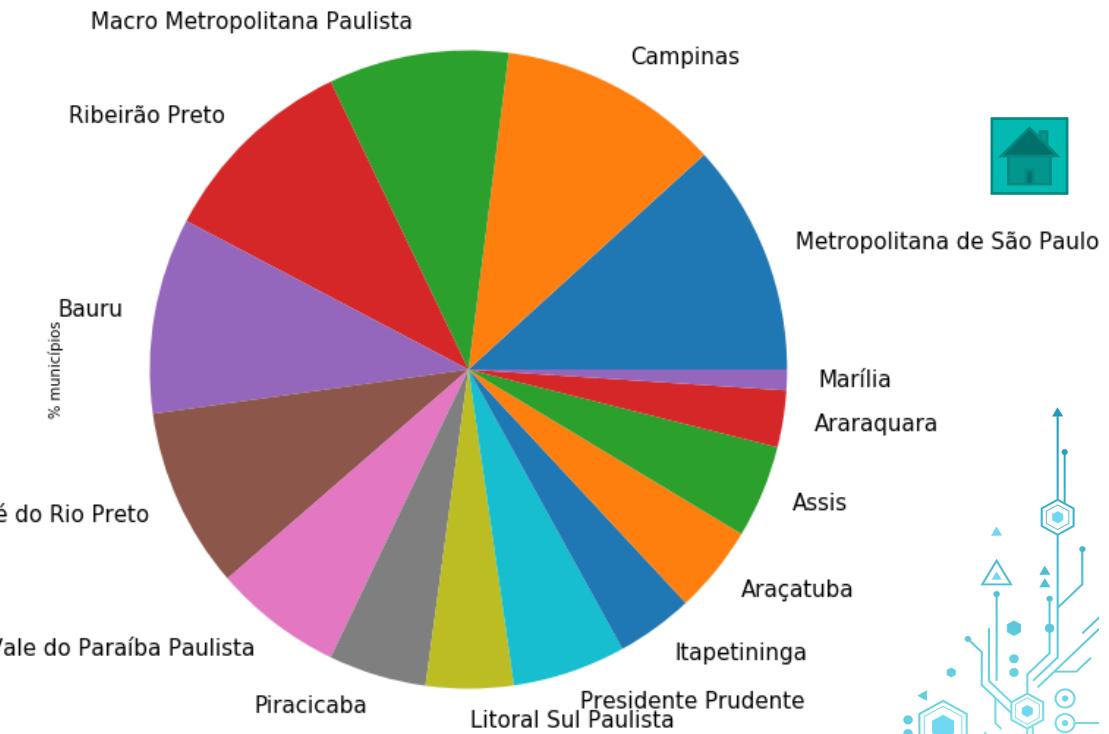
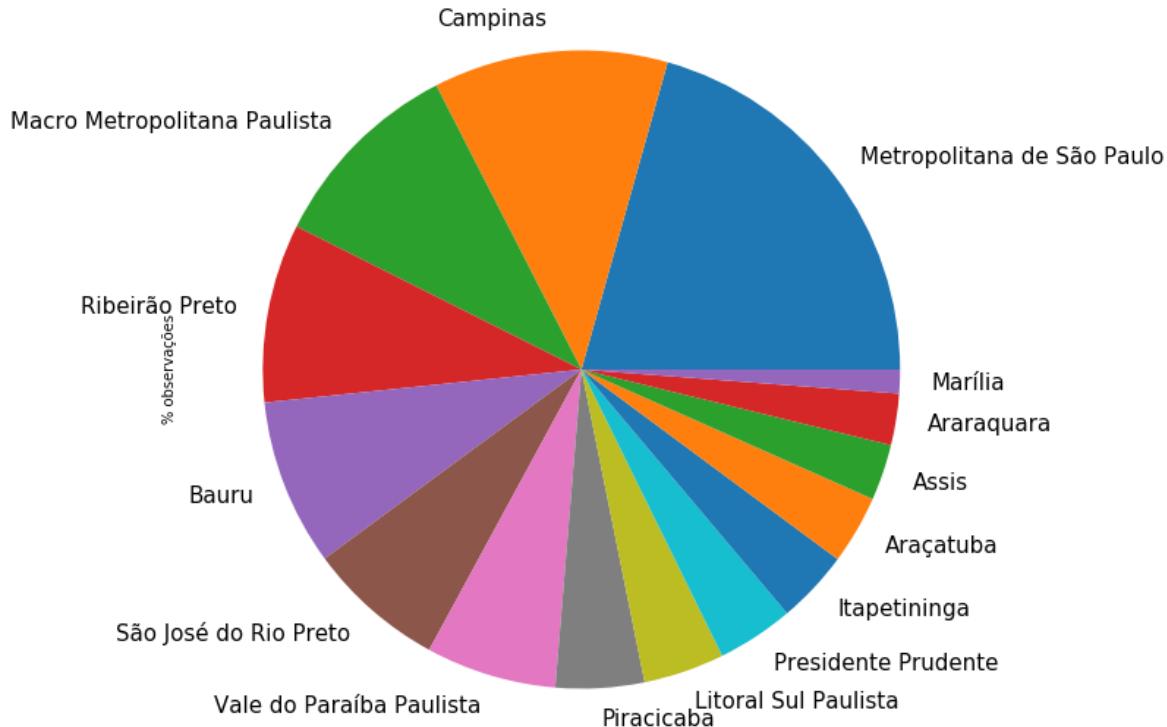
# 4. Análise Exploratória de Dados

RAIO-X DA BASE | ANÁLISE UNIVARIADA

219

## 'Mesorregião Geográfica', 'Nome\_Mesorregião'

- Na análise por mesorregião, se destaca negativamente a mesorregião Metropolitana de São Paulo, com o dobro de % de observações em relação ao % de municípios, demonstrando que foi a mesorregião do estado em que a pandemia mais se proliferou sem controle até a data de corte.
- No outro espectro, destacam-se as mesorregiões de São José do Rio Preto, Presidente Prudente e Assis que, até a data de corte, demonstraram taxas de contágio menores entre municípios.

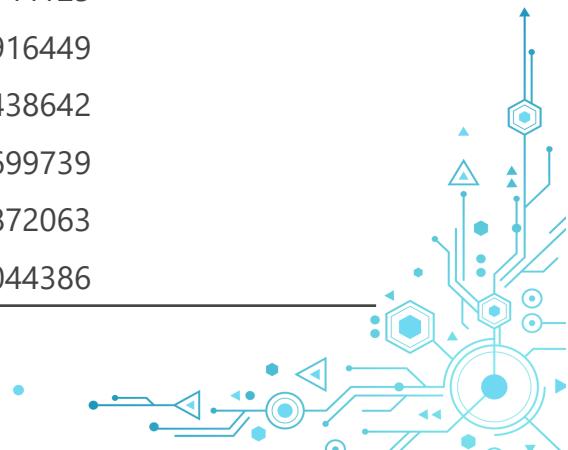


# Detalhes das análises

'MESORREGIÃO' | DISTRIBUIÇÃO DE OBSERVAÇÕES E DE MUNICÍPIOS POR MESORREGIÃO

220

mesorregião	qtd. observações	% observações	qtd. municípios	% municípios
Metropolitana de São Paulo	1815	20.636725	45	11.749347
Campinas	1046	11.893121	43	11.227154
Macro Metropolitana Paulista	886	10.073906	35	9.138381
Ribeirão Preto	796	9.050597	39	10.182768
Bauru	742	8.436612	38	9.921671
São José do Rio Preto	614	6.981239	35	9.138381
Vale do Paraíba Paulista	582	6.617396	25	6.527415
Piracicaba	395	4.491188	19	4.960836
Litoral Sul Paulista	361	4.104605	17	4.438642
Presidente Prudente	341	3.877203	22	5.744125
Itapetininga	326	3.706652	15	3.916449
Araçatuba	306	3.479250	17	4.438642
Assis	250	2.842524	18	4.699739
Araraquara	229	2.603752	11	2.872063
Marília	106	1.205230	4	1.044386



# Detalhes das análises

'MICRORREGIÃO' | DISTRIBUIÇÃO DE OBSERVAÇÕES E DE MUNICÍPIOS POR MICRORREGIÃO

221

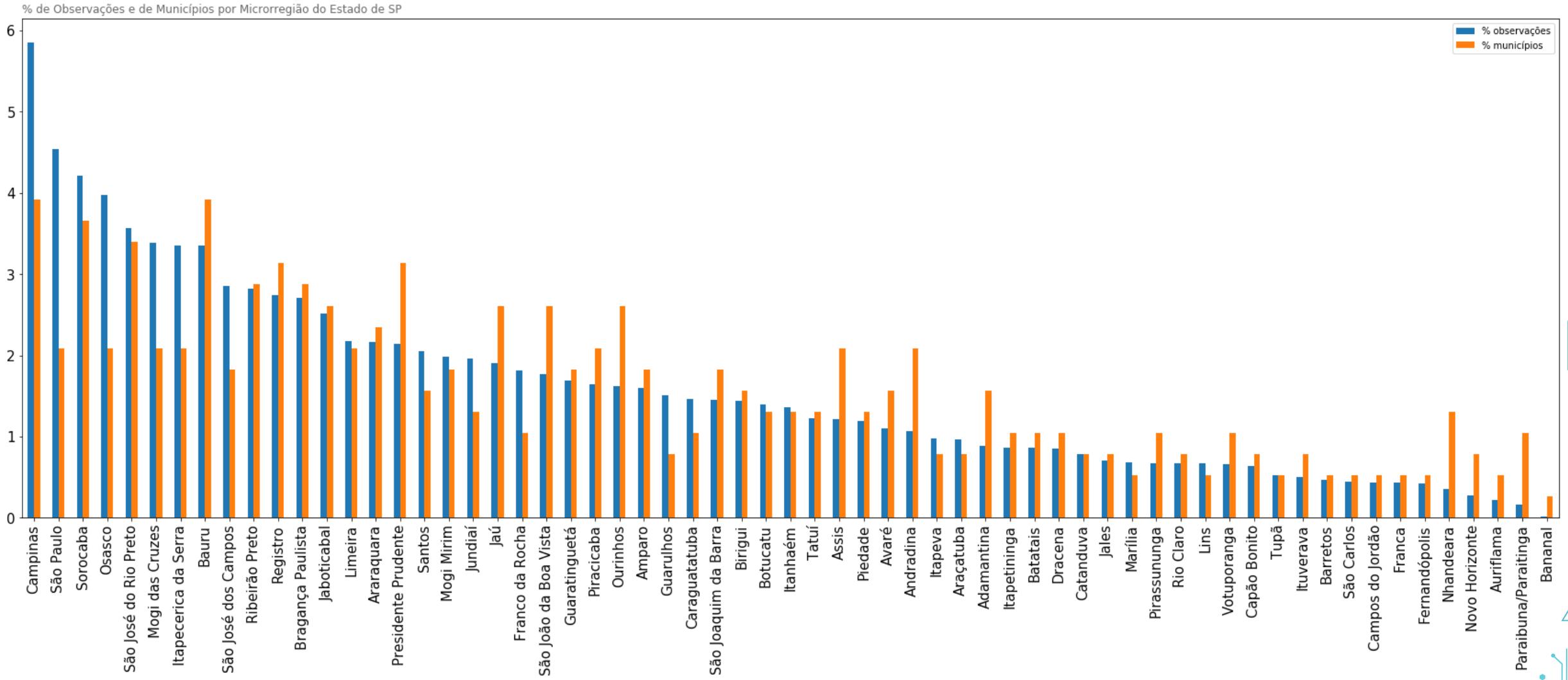
microrregião	qtd. observações	% observações	qtd. municípios	% municípios
Campinas	515	5.855600	15	3.916449
São Paulo	399	4.536669	8	2.088773
Sorocaba	370	4.206936	14	3.655352
Osasco	350	3.979534	8	2.088773
São José do Rio Preto	314	3.570210	13	3.394256
...	...	...	...	...
Nhandeara	31	0.352473	5	1.305483
Novo Horizonte	24	0.272882	3	0.783290
Auriflama	19	0.216032	2	0.522193
Paraibuna/Paraitinga	14	0.159181	4	1.044386
Bananal	1	0.011370	1	0.261097



# Detalhes das análises

'MICRORREGIÃO' | DISTRIBUIÇÃO DE OBSERVAÇÕES E DE MUNICÍPIOS POR MICRORREGIÃO

222



# Detalhes das análises

'DATA' | DISTRIBUIÇÃO DE OBSERVAÇÕES POR DATA

223

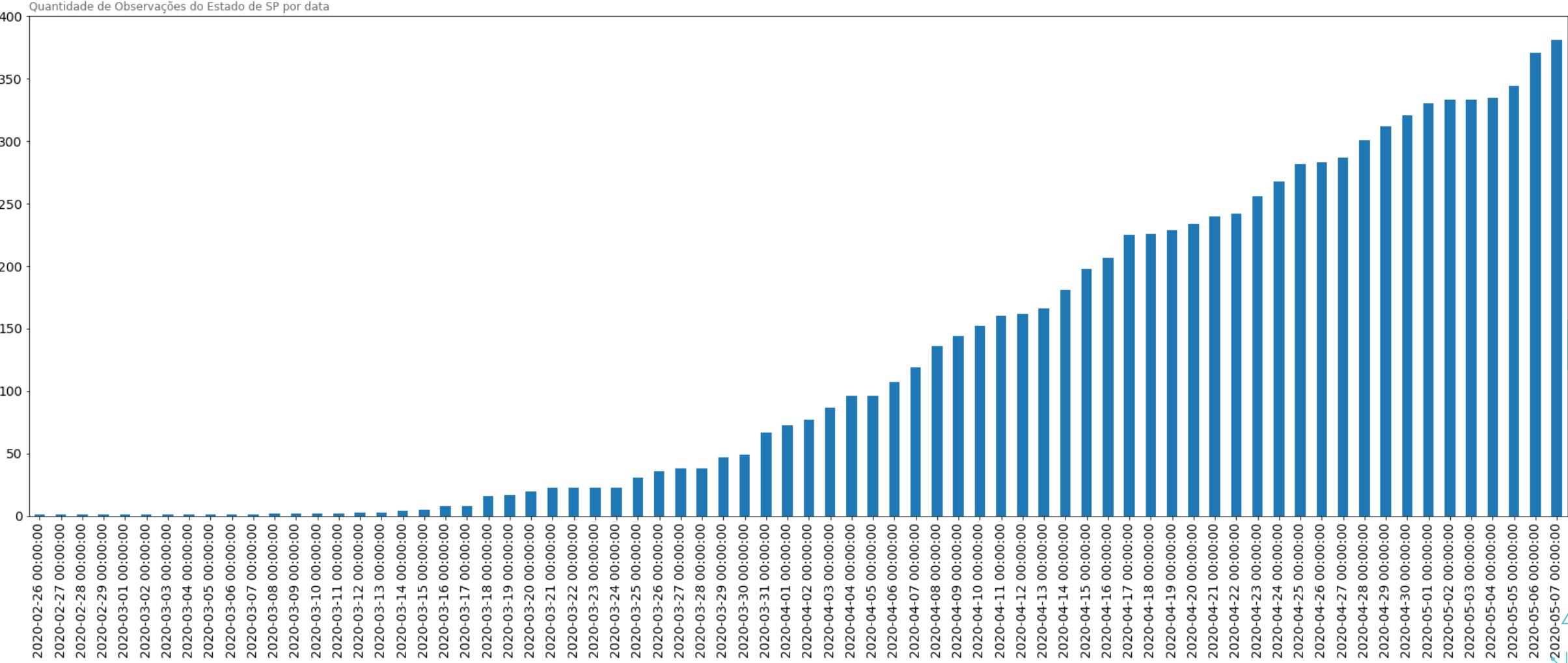
<b>data</b>	<b>qtd.</b>	<b>%</b>
2020-02-26	1	0.011370
2020-02-27	1	0.011370
2020-02-28	1	0.011370
2020-02-29	1	0.011370
2020-03-01	1	0.011370
...	...	...
2020-05-03	333	3.786242
2020-05-04	335	3.808982
2020-05-05	344	3.911313
2020-05-06	371	4.218306
2020-05-07	381	4.332007



# Detalhes das análises

'DATA' | DISTRIBUIÇÃO DE OBSERVAÇÕES POR DATA

224



# Detalhes das análises

'DIAS EPIDEMIOLÓGICOS' | DISTRIBUIÇÃO DE OBSERVAÇÕES POR DIAS EPIDEMIOLÓGICOS

225

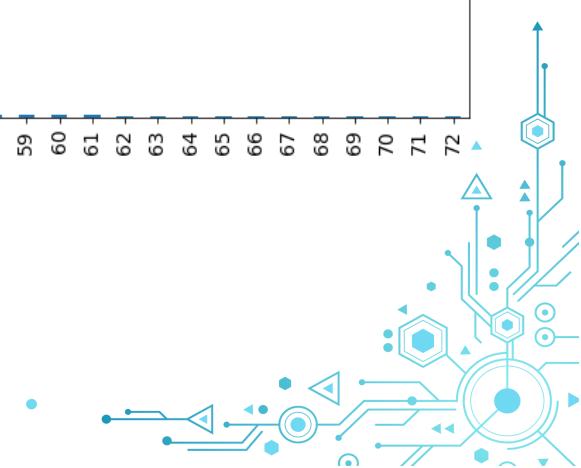
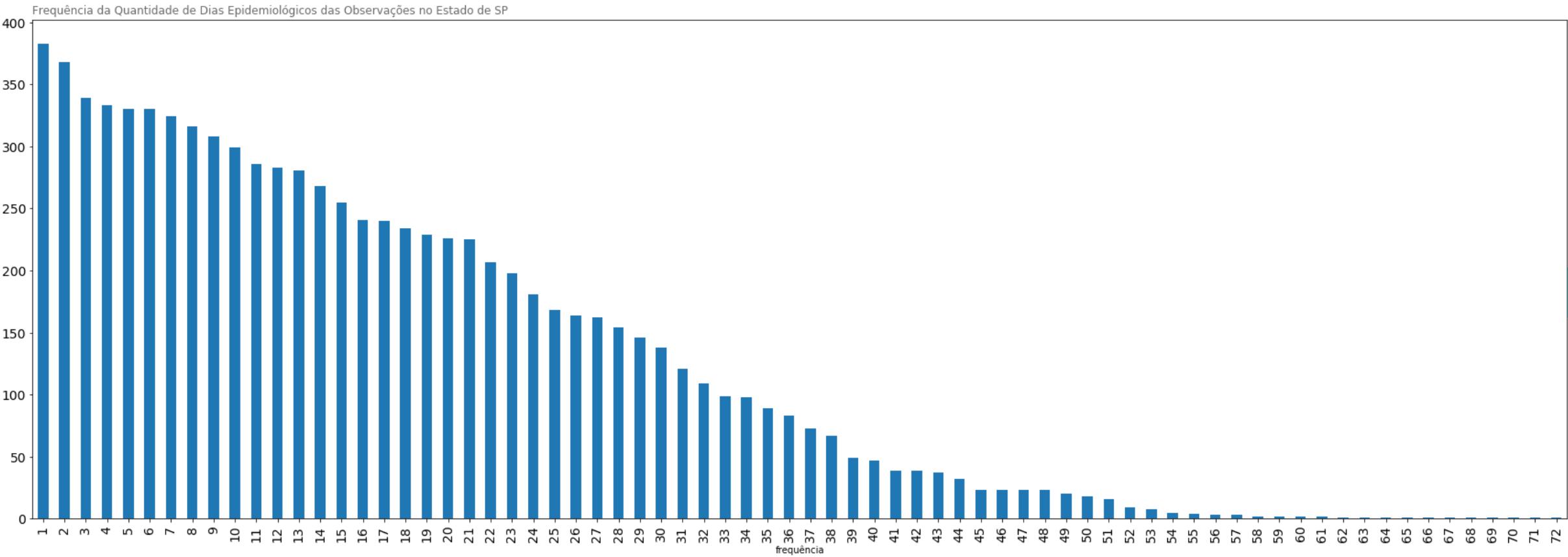
dia epidemiológico	qtd.	%
1	383	4.354747
2	368	4.184196
3	339	3.854463
4	333	3.786242
5	330	3.752132
...	...	...
68	1	0.011370
69	1	0.011370
70	1	0.011370
71	1	0.011370
72	1	0.011370



# Detalhes das análises

'DIAS EPIDEMIOLÓGICOS' | DISTRIBUIÇÃO DE OBSERVAÇÕES POR DIAS EPIDEMIOLÓGICOS

226



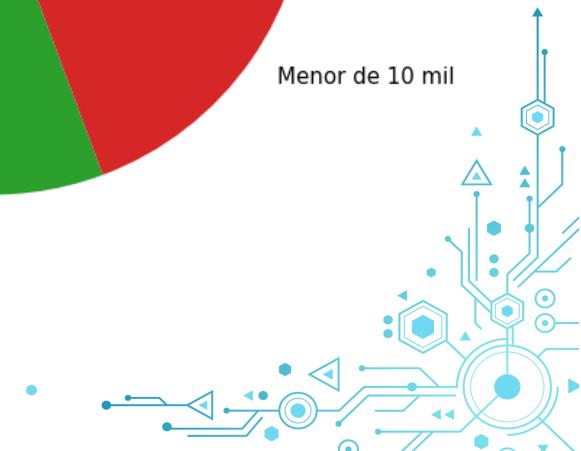
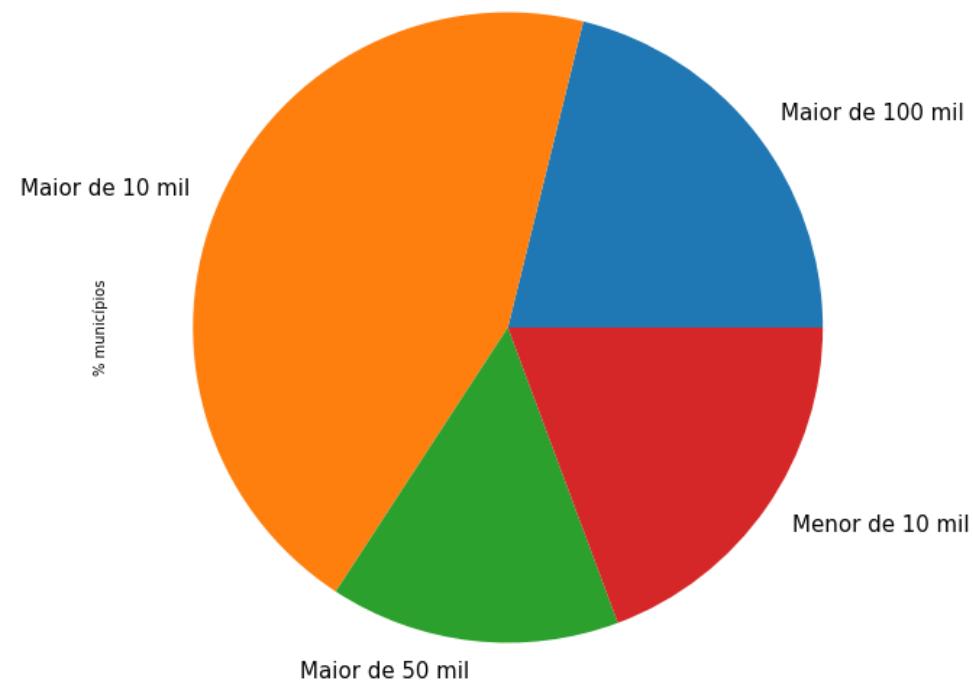
## 4. Análise Exploratória de Dados

RAIO-X DA BASE | ANÁLISE UNIVARIADA

227

'faixa\_pop'

- A maioria das observações (37%) está concentrada nos municípios com população maior de 100 mil habitantes (21% dos municípios) e 1/3 das observações está nas cidades entre 10 mil e 50 mil habitantes (que representam 45% dos municípios).

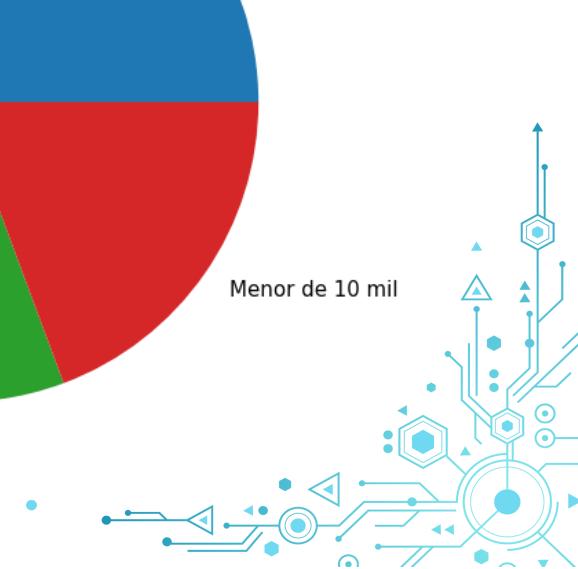
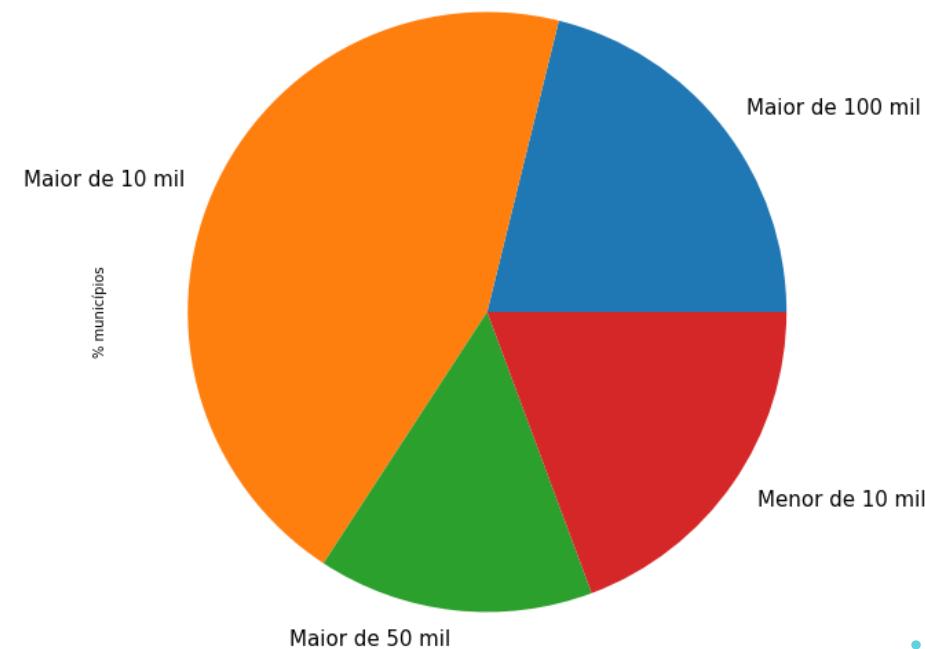
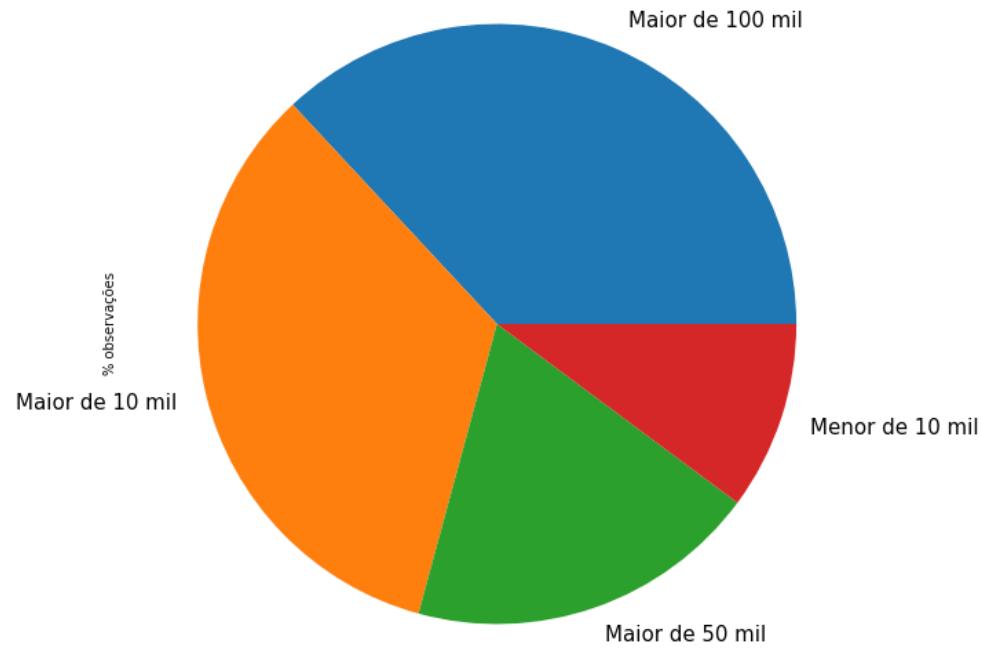


# Detalhes das análises

'FAIXA POPULAÇÃO' | DISTRIBUIÇÃO DE OBSERVAÇÕES E DE MUNICÍPIOS POR FAIXA POPULACIONAL

228

faixa populacional	qtd. observações	% observações	qtd. municípios	% municípios
Maior de 100 mil	3250	36.952814	81	21.148825
Maior de 10 mil	2975	33.826038	171	44.647520
Maior de 50 mil	1677	19.067652	57	14.882507
Menor de 10 mil	893	10.153496	74	19.321149



# 4. Análise Exploratória de Dados

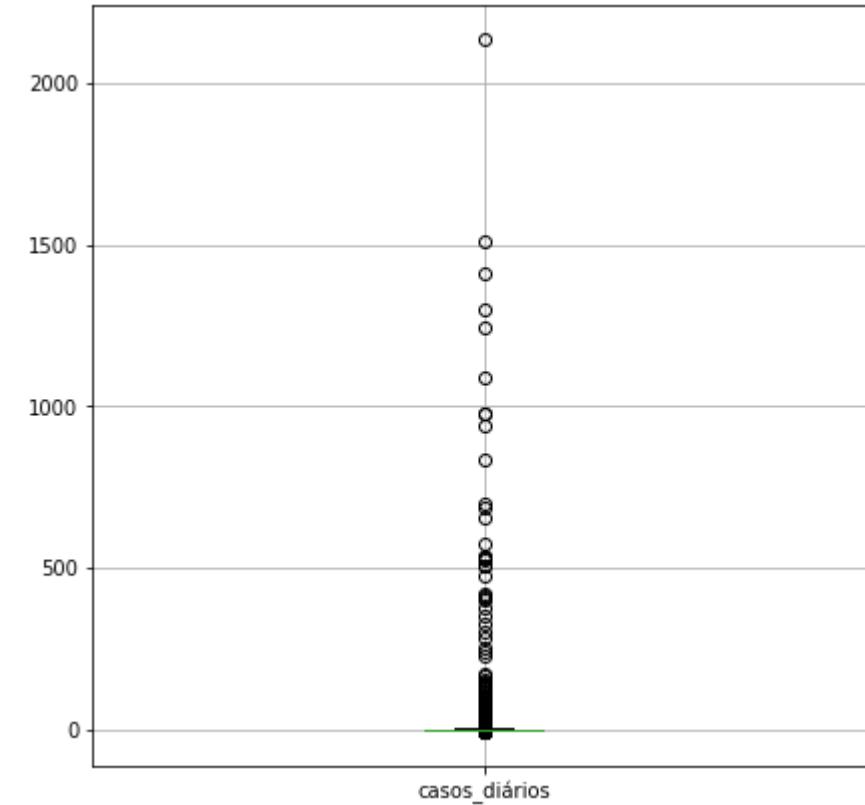
RAIO-X DA BASE | ANÁLISE UNIVARIADA

229

## 'casos diários'

- 50% das observações são de 0 casos diários, o que faz sentido, pois os primeiros dias de transmissão num município costumam ser mais controlados, com baixa taxa de transmissão.
- A moda e a mediana apontarem 0 também confirma o observado no boxplot.
- O 3º quartil é de observações com 1 caso diário.
- A partir daí, os casos diários costumam subir em progressão geométrica, donde que não consideramos outliers os pontos acima do 3º quartil, incluindo o ponto máximo, de 2134 casos diários.
- As poucas observações com casos diários negativos são consideradas normais, devido a flutuações nos registros (casos que são reclassificados ou erros de registro).
- No período analisado, o estado de São Paulo teve quase 40 mil casos diários.
- Os altos valores de amplitude (2142), variância (2514) e coeficiente de variação (1104%) demonstram como os valores de casos diários estão espalhados e distantes da média (4.5).

casos_diários	
Contagem	8795
Média	4.540080
Desvio Padrão	50.142921
Mínimo	-8
25%	0
50%	0
75%	1
Máximo	2134
Soma	39930.0
Moda	0
Mediana	0
Amplitude	2142
Variância	2514.312557
Coeficiente de Variação	1104.45027



# 4. Análise Exploratória de Dados

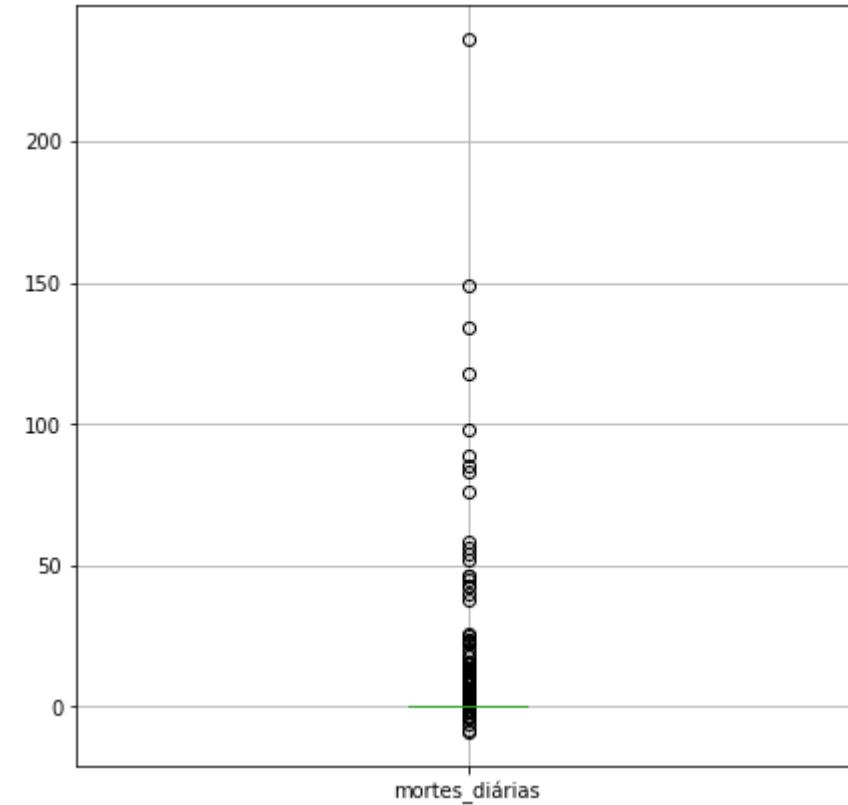
RAIO-X DA BASE | ANÁLISE UNIVARIADA

230

## 'mortes diárias'

- 75% das observações são de 0 mortes diárias, o que faz sentido, pois estima-se que a letalidade da COVID-19 gira em torno de 0.5% dos casos e as mortes diárias só costumam aumentar quando há colapso do sistema de saúde.
- A moda e a mediana apontarem 0 também confirma o observado no boxplot.
- A partir daí, as mortes diárias costumam subir em progressão geométrica, donde que não consideramos outliers os pontos acima do 3º quartil, incluindo o ponto máximo, de 236 mortes diárias.
- As poucas observações com mortes diárias negativos são consideradas normais, devido a flutuações nos registros (casos que são reclassificados ou erros de registro).
- No período analisado, o estado de São Paulo teve quase 3206 mortes.
- Os valores de amplitude (245) e variância (20.9) não são tão altos quanto os dos casos diários, mas o coeficiente de variação (1254%) está na mesma escala dos casos diários, demonstrando como os valores de mortes diárias estão espalhados e distantes da média (0.36).

mortes diárias	
Contagem	8795
Média	0.364525
Desvio Padrão	4.572631
Mínimo	-9
25%	0
50%	0
75%	0
Máximo	236
Soma	3206.0
Moda	0
Mediana	0
Amplitude	245.0
Variância	20.908953
Coeficiente de Variação	1254.406998



# 4. Análise Exploratória de Dados

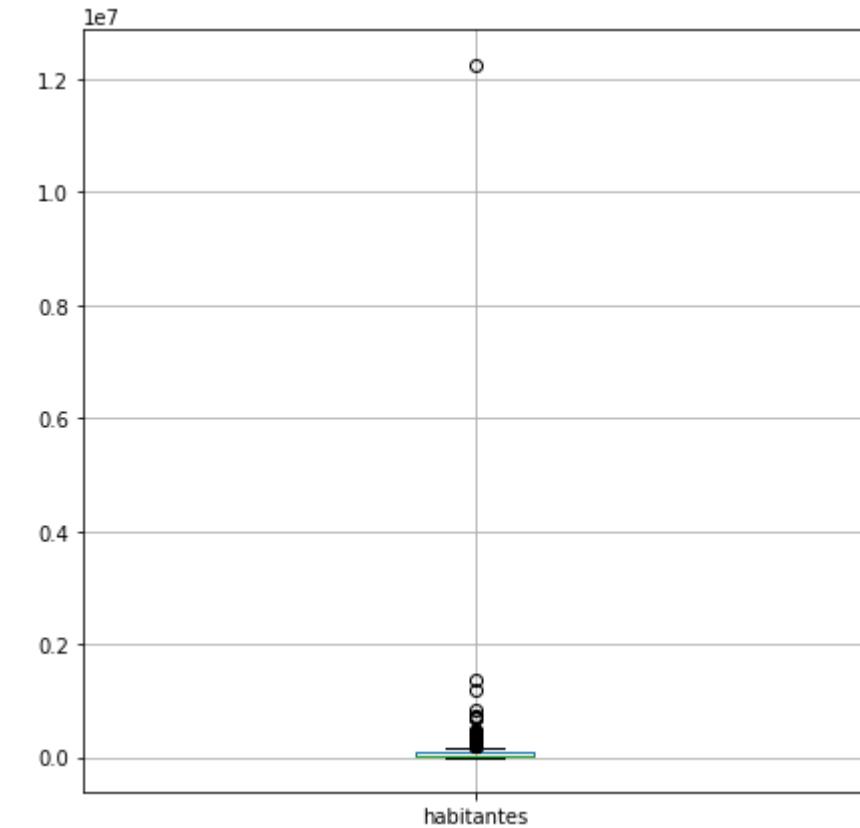
RAIO-X DA BASE | ANÁLISE UNIVARIADA

231

## 'habitantes'

- Dos 383 municípios paulistas com ao menos 1 caso de COVID-19, 25% têm até 12870 habitantes.
- Dos 383 municípios paulistas com ao menos 1 caso de COVID-19, 50% têm até 30857 habitantes.
- Dos 383 municípios paulistas com ao menos 1 caso de COVID-19, 75% têm até 78803 habitantes.
- Os 383 municípios paulistas com ao menos 1 caso de COVID-19 possuem ao todo 43,711,457 habitantes, dos quais 12,252,023 estão na capital São Paulo.
- Os altos valores de amplitude (12,250,858), variância (409,239,292,578) e coeficiente de variação (560%) demonstram como o nº de habitantes nos municípios está espalhado e distante da média (114,129).

habitantes	
Contagem	383
Média	114129.13055
Desvio Padrão	639718.13526
Mínimo	1165
25%	12870
50%	30857
75%	78803
Máximo	12252023
Soma	43711457
Moda	6929
Mediana	30857
Amplitude	12250858
Variância	409239292578 .42285
Coeficiente de Variação	560.52134



# 4. Análise Exploratória de Dados

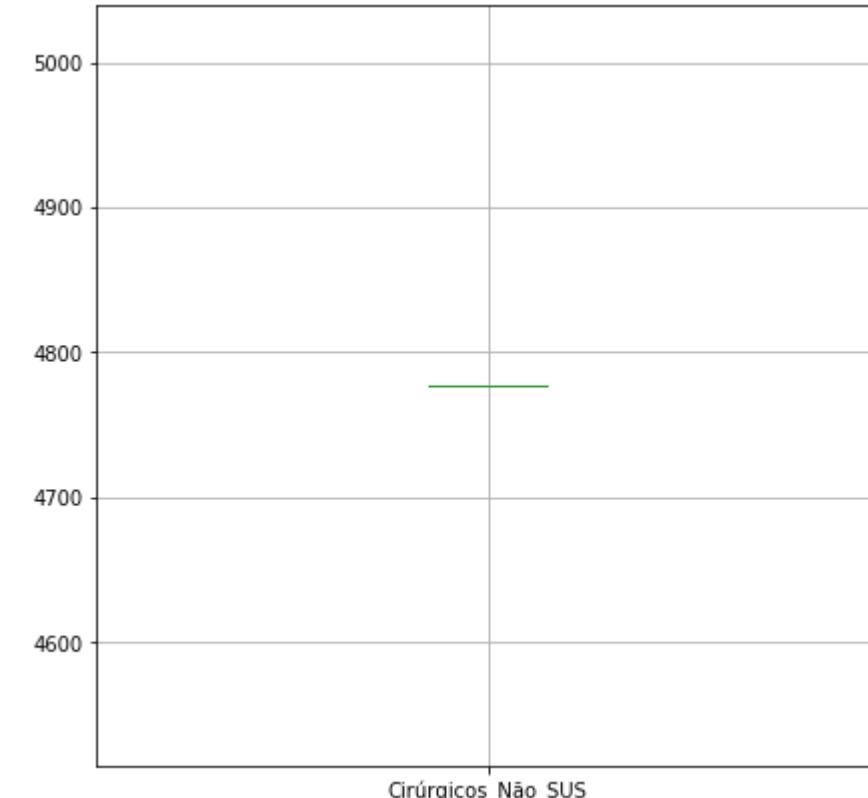
RAIO-X DA BASE | ANÁLISE UNIVARIADA

232

## 'Leitos Cirúrgicos Não SUS' – mês 2

- Obs.: não faz sentido fazer análise univariada de nº de leitos hospitalares, pois eles são únicos para cada cidade e para cada mês. Portanto, vamos analisar o nº de leitos por mês.
- No mês 2, só havia observações da cidade de São Paulo, que apresentava 4777 leitos cirúrgicos particulares.

	Leitos Cirúrgicos Não SUS
Contagem	1
Média	4777
Desvio Padrão	-
Mínimo	4777
25%	4777
50%	4777
75%	4777
Máximo	4777
Soma	4777
Moda	4777
Mediana	4777
Amplitude	0
Variância	-
Coeficiente de Variação	-



# 4. Análise Exploratória de Dados

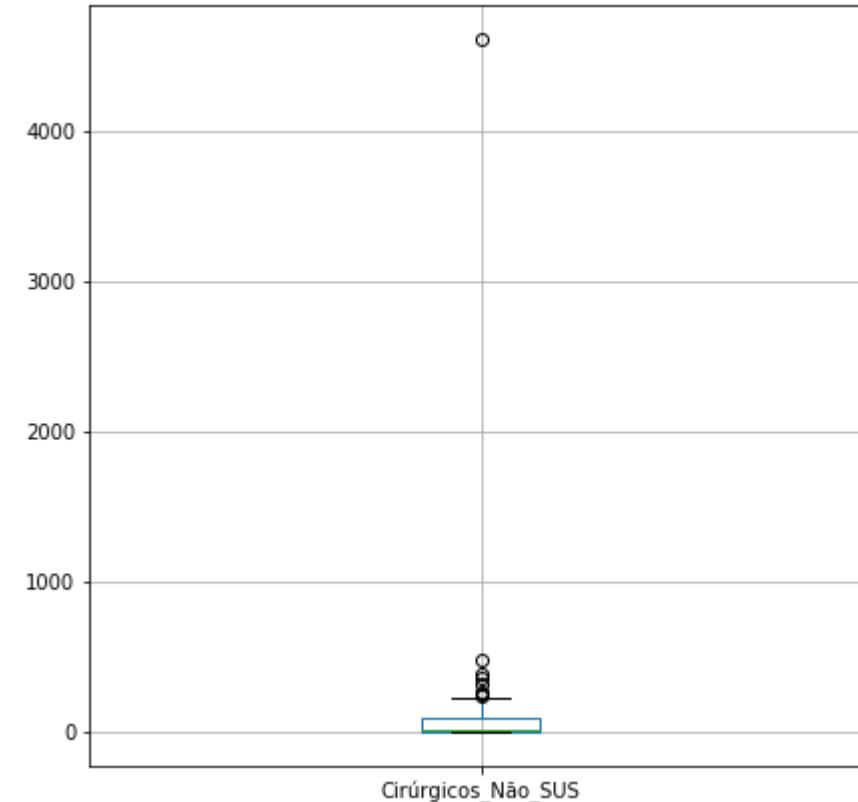
RAIO-X DA BASE | ANÁLISE UNIVARIADA

233

## 'Leitos Cirúrgicos Não SUS' – mês 3

- No mês 3, havia 67 municípios do estado de São Paulo com ao menos 1 caso de COVID-19. Destes, a situação de leitos cirúrgicos particulares foi a seguinte:
  - O 1º quartil dos 67 municípios não possuía leitos cirúrgicos particulares.
  - O 2º quartil dos 67 municípios possuía até 18 leitos cirúrgicos particulares.
  - O 3º quartil dos 67 municípios possuía até 96 leitos cirúrgicos particulares.
  - A cidade com maior oferta de leitos cirúrgicos particulares foi São Paulo, com 4611 (o outlier no boxplot, que representa mais do que a soma de todos os outros 66 municípios), o que denota uma queda em relação ao mês anterior.
  - Os 67 municípios paulistas com ao menos 1 caso de COVID-19 no mês 3 possuíam ao todo 9103 leitos cirúrgicos particulares.
  - Os altos valores de amplitude (4611), variância (319798) e coeficiente de variação (416%) demonstram como o nº de leitos cirúrgicos particulares nos municípios está espalhado e distante da média (135).

Leitos Cirúrgicos Não SUS	
Contagem	67
Média	135.86567
Desvio Padrão	565.50698
Mínimo	0
25%	0.5
50%	18
75%	96
Máximo	4611
Soma	9103
Moda	0
Mediana	18
Amplitude	4611
Variância	319798.14835
Coeficiente de Variação	416.22507



# 4. Análise Exploratória de Dados

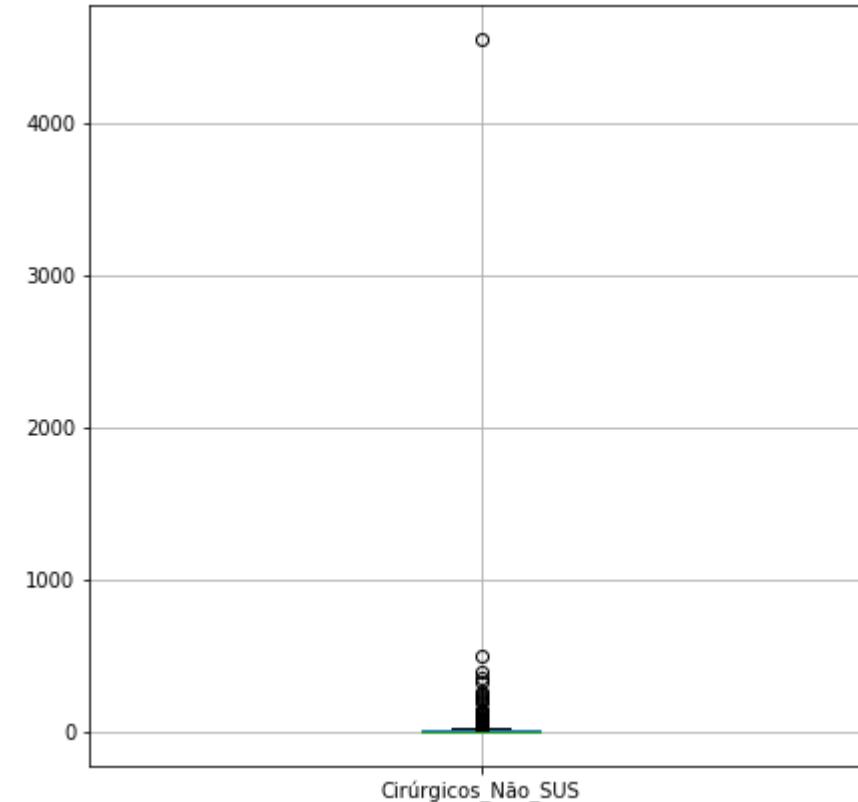
RAIO-X DA BASE | ANÁLISE UNIVARIADA

234

## 'Leitos Cirúrgicos Não SUS' – mês 4

- No mês 4, havia 321 municípios do estado de São Paulo com ao menos 1 caso de COVID-19. Destes, a situação de leitos cirúrgicos particulares foi a seguinte:
  - O 1º quartil dos 321 municípios não possuía leitos cirúrgicos particulares.
  - O 2º quartil dos 321 municípios possuía 1 leito cirúrgico particular.
  - O 3º quartil dos 321 municípios possuía até 14 leitos cirúrgicos particulares.
  - A cidade com maior oferta de leitos cirúrgicos particulares foi São Paulo, com 4546 (o outlier no boxplot, que representa 2/3 da soma de todos os outros 320 municípios), o que denota novamente uma queda em relação ao mês anterior.
  - Os 321 municípios paulistas com ao menos 1 caso de COVID-19 no mês 4 possuíam ao todo 11362 leitos cirúrgicos particulares.
  - Os altos valores de amplitude (4546), variância (67108) e coeficiente de variação (731%) demonstram como o nº de leitos cirúrgicos particulares nos municípios está espalhado e distante da média (35).

Leitos Cirúrgicos Não SUS	
Contagem	321
Média	35.39564
Desvio Padrão	259.05321
Mínimo	0
25%	0
50%	1
75%	14
Máximo	4546
Soma	11362
Moda	0
Mediana	1
Amplitude	4546
Variância	67108.56486
Coeficiente de Variação	731.87889



# 4. Análise Exploratória de Dados

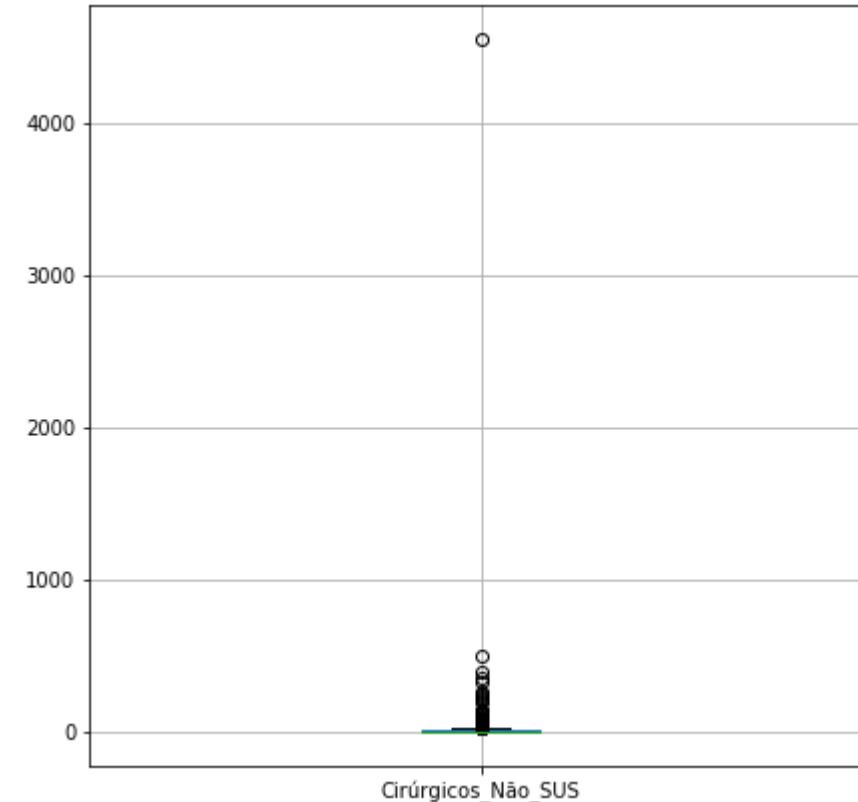
RAIO-X DA BASE | ANÁLISE UNIVARIADA

235

## 'Leitos Cirúrgicos Não SUS' – mês 5

- No mês 5, havia 383 municípios do estado de São Paulo com ao menos 1 caso de COVID-19. Destes, a situação de leitos cirúrgicos particulares foi a seguinte:
  - O 1º quartil dos 383 municípios não possuía leitos cirúrgicos particulares.
  - O 2º quartil dos 383 municípios possuía 1 leito cirúrgico particular.
  - O 3º quartil dos 383 municípios possuía até 11 leitos cirúrgicos particulares.
  - A cidade com maior oferta de leitos cirúrgicos particulares foi São Paulo, com 4546 (o outlier no boxplot, que representa 2/3 da soma de todos os outros 382 municípios), ou seja, São Paulo manteve o mesmo nº de leitos cirúrgicos particulares em relação ao mês anterior.
  - Os 383 municípios paulistas com ao menos 1 caso de COVID-19 no mês 5 possuíam ao todo 11416 leitos cirúrgicos particulares.
  - Os altos valores de amplitude (4546), variância (56379) e coeficiente de variação (796%) demonstram como o nº de leitos cirúrgicos particulares nos municípios está espalhado e distante da média (29).

Leitos Cirúrgicos Não SUS	
Contagem	383
Média	29.80679
Desvio Padrão	237.44363
Mínimo	0
25%	0
50%	1
75%	11
Máximo	4546
Soma	11416
Moda	0
Mediana	1
Amplitude	4546
Variância	56379.47566
Coeficiente de Variação	796.60922



# 4. Análise Exploratória de Dados

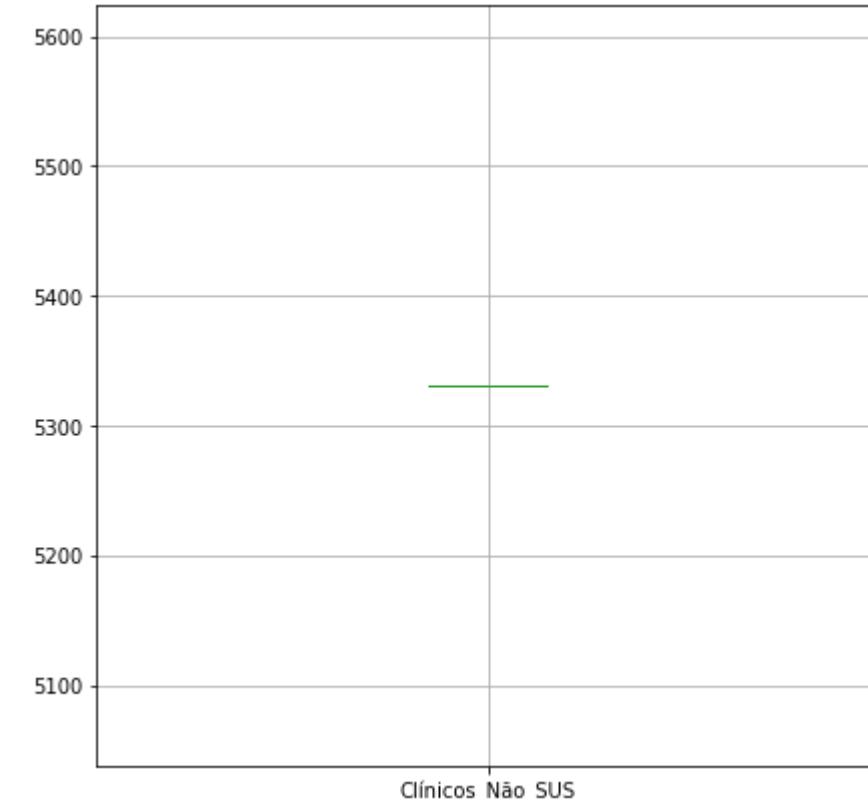
RAIO-X DA BASE | ANÁLISE UNIVARIADA

236

## 'Leitos Clínicos Não SUS' – mês 2

- Obs.: não faz sentido fazer análise univariada de nº de leitos hospitalares, pois eles são únicos para cada cidade e para cada mês. Portanto, vamos analisar o nº de leitos por mês.
- No mês 2, só havia observações da cidade de São Paulo, que apresentava 5331 leitos clínicos particulares.

Leitos Clínicos Não SUS	
Contagem	1
Média	5331
Desvio Padrão	-
Mínimo	5331
25%	5331
50%	5331
75%	5331
Máximo	5331
Soma	5331
Moda	5331
Mediana	5331
Amplitude	0
Variância	-
Coeficiente de Variação	-



# 4. Análise Exploratória de Dados

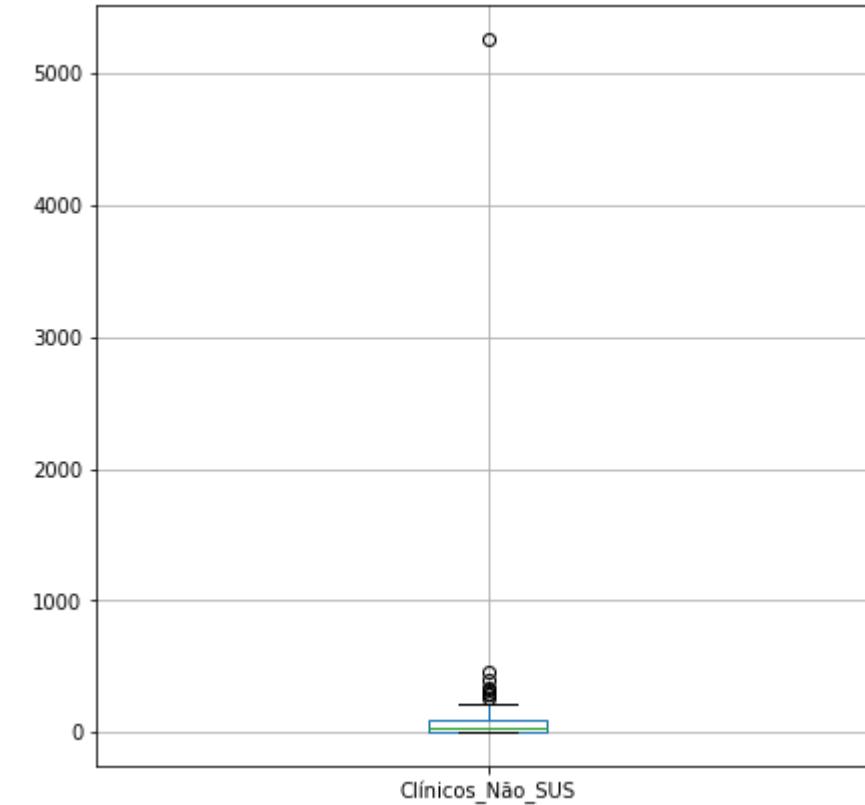
RAIO-X DA BASE | ANÁLISE UNIVARIADA

237

## 'Leitos Clínicos Não SUS' - mês 3

- No mês 3, havia 67 municípios do estado de São Paulo com ao menos 1 caso de COVID-19. Destes, a situação de leitos clínicos particulares foi a seguinte:
  - O 1º quartil dos 67 municípios possuía até 2 leitos clínicos particulares.
  - O 2º quartil dos 67 municípios possuía até 28 leitos clínicos particulares.
  - O 3º quartil dos 67 municípios possuía até 88 leitos clínicos particulares.
  - A cidade com maior oferta de leitos clínicos particulares foi São Paulo, com 5258 (o outlier no boxplot, que representa mais do que a soma de todos os outros 66 municípios), o que denota uma queda em relação ao mês anterior.
  - Os 67 municípios paulistas com ao menos 1 caso de COVID-19 no mês 3 possuíam ao todo 10019 leitos clínicos particulares.
  - Os altos valores de amplitude (5258), variância (413161) e coeficiente de variação (429%) demonstram como o nº de leitos clínicos particulares nos municípios está espalhado e distante da média (149).

Leitos Clínicos Não SUS	
Contagem	67
Média	149.53731
Desvio Padrão	642.77611
Mínimo	0
25%	2.5
50%	28
75%	88.5
Máximo	5258
Soma	10019
Moda	0
Mediana	28
Amplitude	5258
Variância	413161.13116
Coeficiente de Variação	429.84329



# 4. Análise Exploratória de Dados

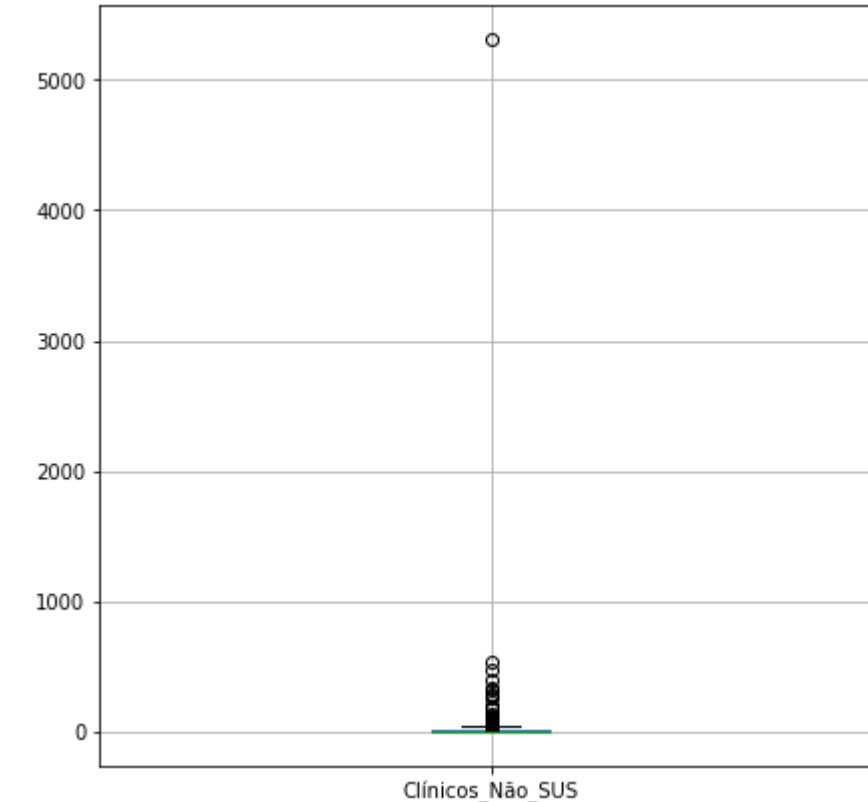
RAIO-X DA BASE | ANÁLISE UNIVARIADA

238

## 'Leitos Clínicos Não SUS' – mês 4

- No mês 4, havia 321 municípios do estado de São Paulo com ao menos 1 caso de COVID-19. Destes, a situação de leitos clínicos particulares foi a seguinte:
  - O 1º quartil dos 321 municípios não possuía leitos clínicos particulares.
  - O 2º quartil dos 321 municípios possuía até 5 leitos clínicos particulares.
  - O 3º quartil dos 321 municípios possuía até 19 leitos clínicos particulares.
  - A cidade com maior oferta de leitos clínicos particulares foi São Paulo, com 5309 (o outlier no boxplot, que representa quase 2/3 da soma de todos os outros 320 municípios), o que denota um aumento em relação ao mês anterior.
  - Os 321 municípios paulistas com ao menos 1 caso de COVID-19 no mês 4 possuíam ao todo 13363 leitos clínicos particulares.
  - Os altos valores de amplitude (5309), variância (91036) e coeficiente de variação (724%) demonstram como o nº de leitos clínicos particulares nos municípios está espalhado e distante da média (41).

Leitos Clínicos Não SUS	
Contagem	321
Média	41.62928
Desvio Padrão	301.72178
Mínimo	0
25%	0
50%	5
75%	19
Máximo	5309
Soma	13363
Moda	0
Mediana	5
Amplitude	5309
Variância	91036.03401
Coeficiente de Variação	724.78255



# 4. Análise Exploratória de Dados

RAIO-X DA BASE | ANÁLISE UNIVARIADA

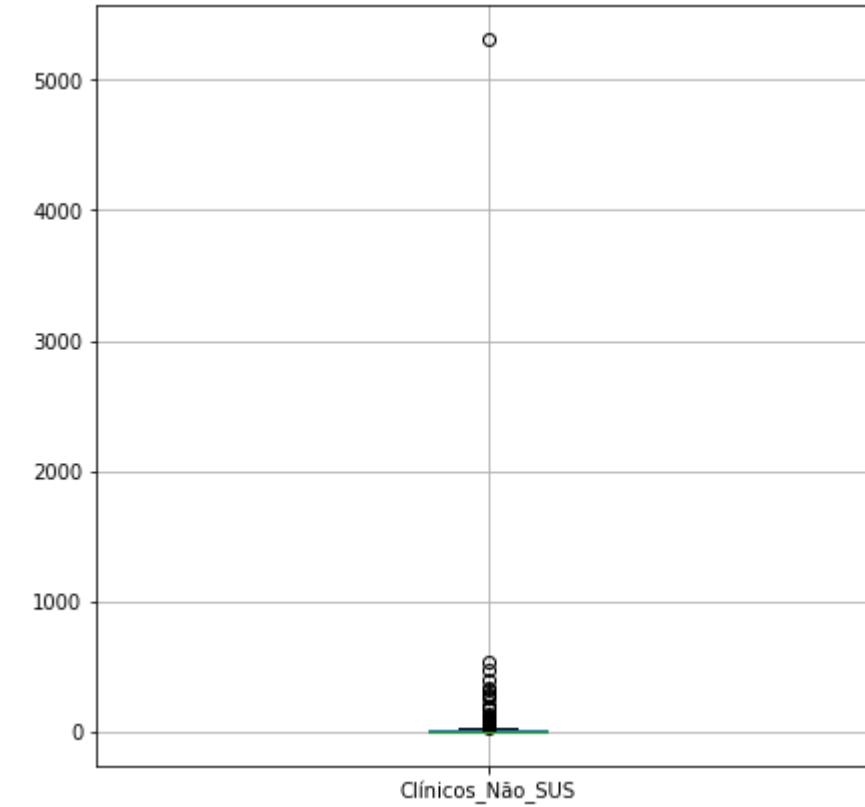
239

## 'Leitos Clínicos Não SUS'

### - mês 5

- No mês 5, havia 383 municípios do estado de São Paulo com ao menos 1 caso de COVID-19. Destes, a situação de leitos clínicos particulares foi a seguinte:
  - O 1º quartil dos 383 municípios não possuía leitos clínicos particulares.
  - O 2º quartil dos 383 municípios possuía até 3 leitos clínicos particulares.
  - O 3º quartil dos 383 municípios possuía até 15 leitos clínicos particulares.
  - A cidade com maior oferta de leitos clínicos particulares foi São Paulo, com 5309 (o outlier no boxplot, que representa quase 2/3 da soma de todos os outros 382 municípios), ou seja, São Paulo manteve o mesmo nº de leitos clínicos particulares em relação ao mês anterior.
  - Os 383 municípios paulistas com ao menos 1 caso de COVID-19 no mês 5 possuíam ao todo 13540 leitos clínicos particulares.
  - Os altos valores de amplitude (5309), variância (76471) e coeficiente de variação (782%) demonstram como o nº de leitos clínicos particulares nos municípios está espalhado e distante da média (35).

Leitos Clínicos Não SUS	
Contagem	383
Média	35.35248
Desvio Padrão	276.53447
Mínimo	0
25%	0
50%	3
75%	15.5
Máximo	5309
Soma	13540
Moda	0
Mediana	3
Amplitude	5309
Variância	76471.31261
Coeficiente de Variação	782.22084



# 4. Análise Exploratória de Dados

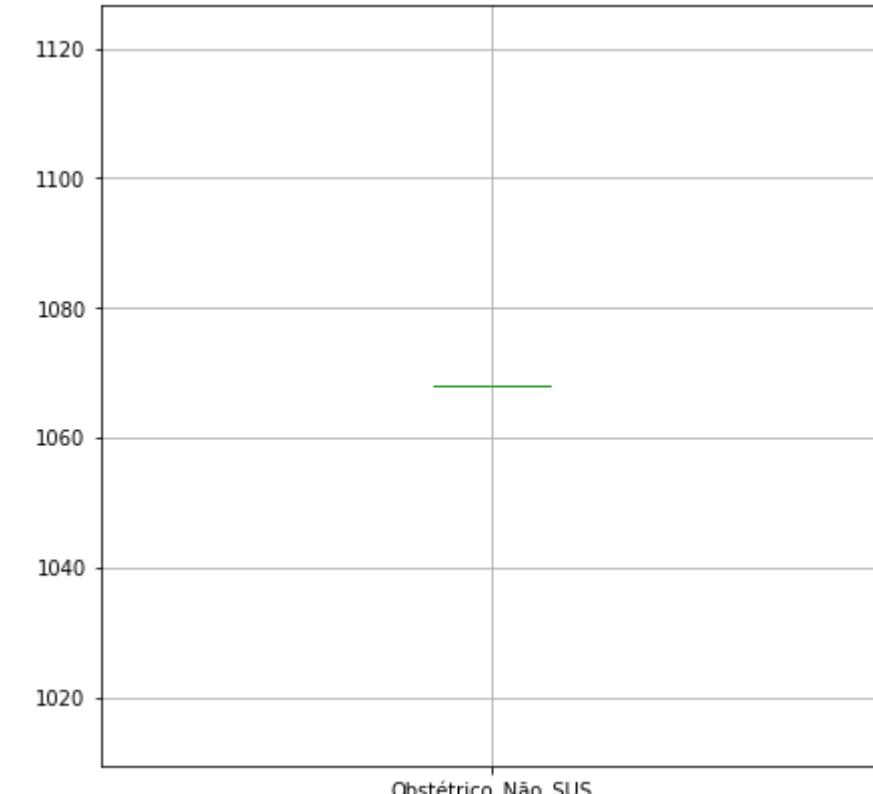
RAIO-X DA BASE | ANÁLISE UNIVARIADA

240

## 'Leitos Obstétricos Não SUS' – mês 2

- Obs.: não faz sentido fazer análise univariada de nº de leitos hospitalares, pois eles são únicos para cada cidade e para cada mês. Portanto, vamos analisar o nº de leitos por mês.
- No mês 2, só havia observações da cidade de São Paulo, que apresentava 1068 leitos obstétricos particulares.

	Leitos Obstétricos Não SUS
Contagem	1
Média	1068
Desvio Padrão	-
Mínimo	1068
25%	1068
50%	1068
75%	1068
Máximo	1068
Soma	1068
Moda	1068
Mediana	1068
Amplitude	0
Variância	-
Coeficiente de Variação	-



# 4. Análise Exploratória de Dados

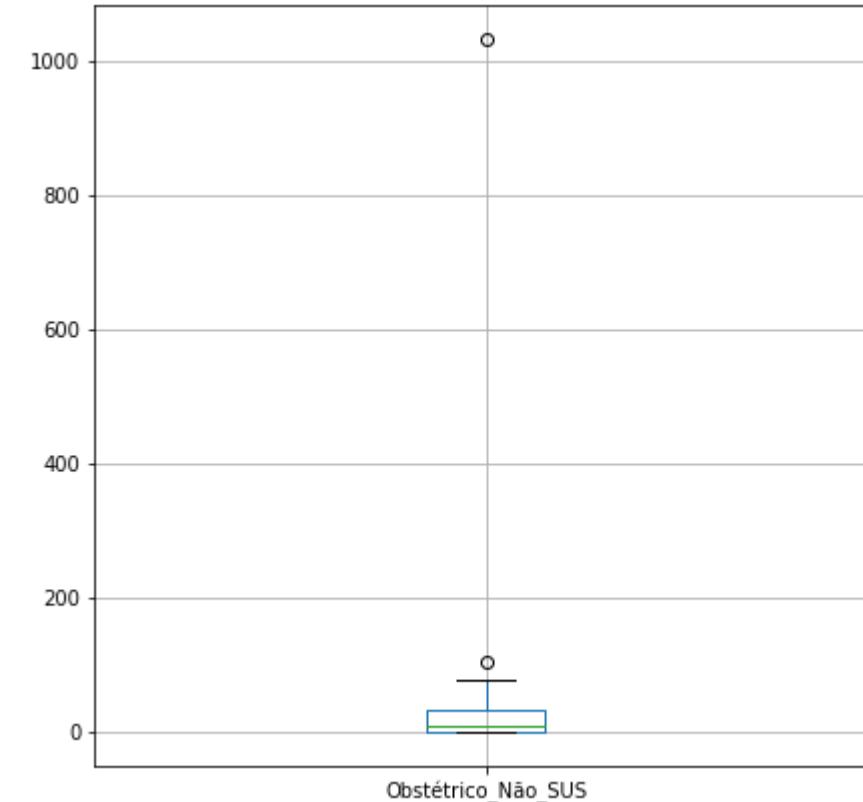
RAIO-X DA BASE | ANÁLISE UNIVARIADA

241

## 'Leitos Obstétricos Não SUS' – mês 3

- No mês 3, havia 67 municípios do estado de São Paulo com ao menos 1 caso de COVID-19. Destes, a situação de leitos obstétricos particulares foi a seguinte:
  - O 1º quartil dos 67 municípios não possuía leitos obstétricos particulares.
  - O 2º quartil dos 67 municípios possuía até 8 leitos obstétricos particulares.
  - O 3º quartil dos 67 municípios possuía até 33 leitos obstétricos particulares.
  - A cidade com maior oferta de leitos obstétricos particulares foi São Paulo, com 1031 (o outlier no boxplot, que representa quase 80% da soma de todos os outros 66 municípios), o que denota uma queda em relação ao mês anterior.
  - Os 67 municípios paulistas com ao menos 1 caso de COVID-19 no mês 3 possuíam ao todo 2333 leitos obstétricos particulares.
  - Os altos valores de amplitude (1031), variância (15927) e coeficiente de variação (362%) demonstram como o nº de leitos obstétricos particulares nos municípios está espalhado e distante da média (34).

Leitos Obstétricos Não SUS	
Contagem	67
Média	34.82090
Desvio Padrão	126.20461
Mínimo	0
25%	0
50%	8
75%	33.5
Máximo	1031
Soma	2333
Moda	0
Mediana	8
Amplitude	1031
Variância	15927.60380
Coeficiente de Variação	362.43930



# 4. Análise Exploratória de Dados

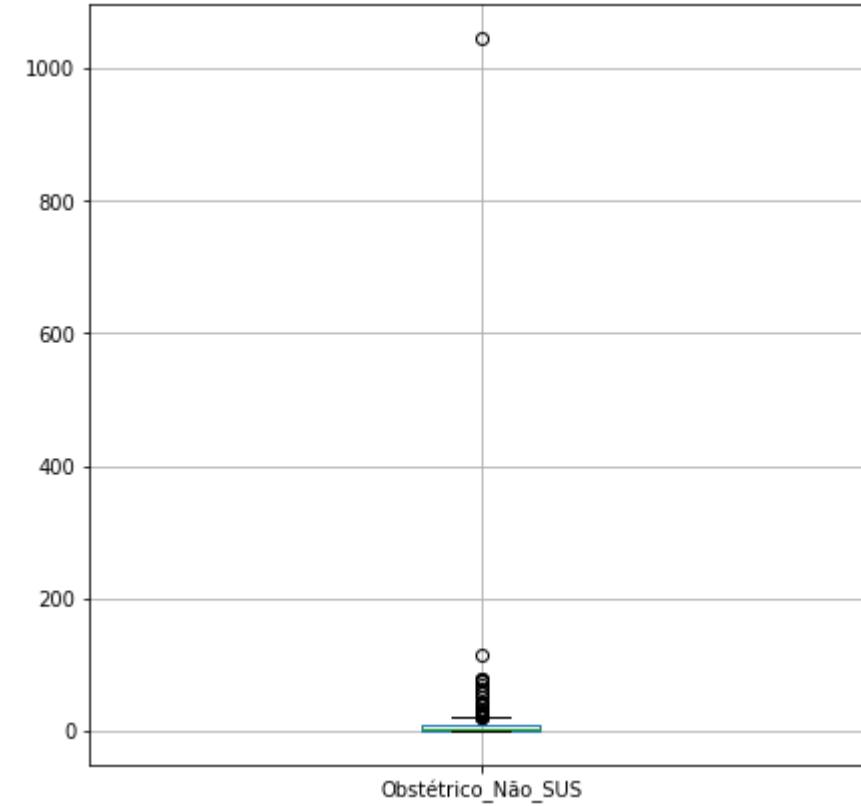
RAIO-X DA BASE | ANÁLISE UNIVARIADA

242

## 'Leitos Obstétricos Não SUS' – mês 4

- No mês 4, havia 321 municípios do estado de São Paulo com ao menos 1 caso de COVID-19. Destes, a situação de leitos obstétricos particulares foi a seguinte:
  - O 1º quartil dos 321 municípios não possuía leitos obstétricos particulares.
  - O 2º quartil dos 321 municípios possuía até 2 leitos obstétricos particulares.
  - O 3º quartil dos 321 municípios possuía até 8 leitos obstétricos particulares.
  - A cidade com maior oferta de leitos obstétricos particulares foi São Paulo, com 1045 (o outlier no boxplot, que representa 45% da soma de todos os outros 320 municípios), o que denota um aumento em relação ao mês anterior.
  - Os 321 municípios paulistas com ao menos 1 caso de COVID-19 no mês 4 possuíam ao todo 3345 leitos obstétricos particulares.
  - Os altos valores de amplitude (1045), variância (3568) e coeficiente de variação (573%) demonstram como o nº de leitos obstétricos particulares nos municípios está espalhado e distante da média (10).

Leitos Obstétricos Não SUS	
Contagem	321
Média	10.42056
Desvio Padrão	59.73614
Mínimo	0
25%	0
50%	2
75%	8
Máximo	1045
Soma	3345
Moda	0
Mediana	2
Amplitude	1045
Variância	3568.40695
Coeficiente de Variação	573.25269



# 4. Análise Exploratória de Dados

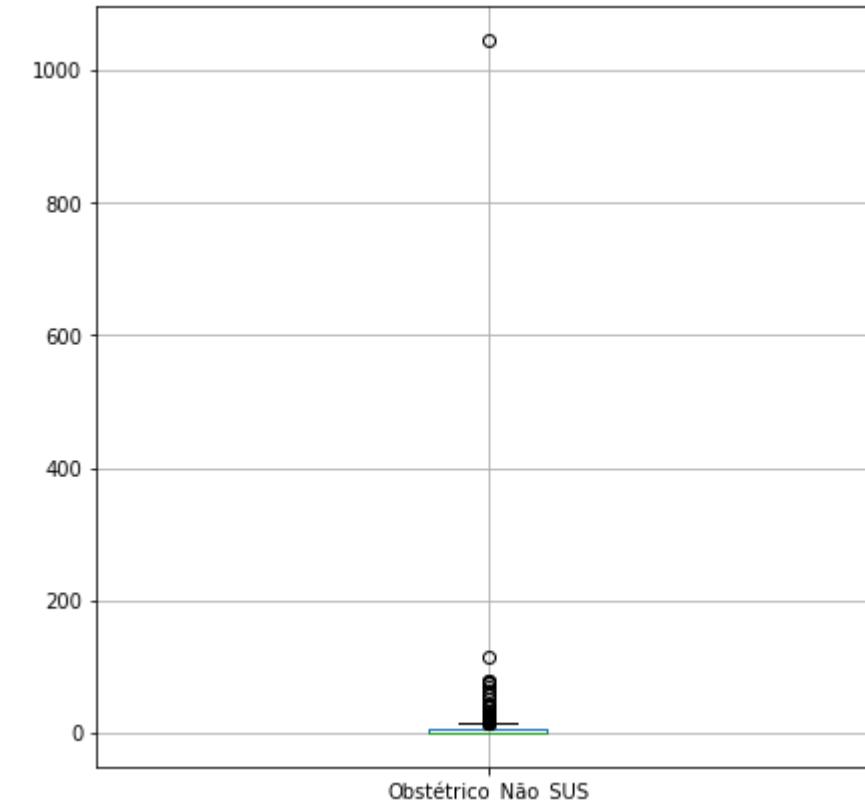
RAIO-X DA BASE | ANÁLISE UNIVARIADA

243

## 'Leitos Obstétricos Não SUS' – mês 5

- No mês 5, havia 383 municípios do estado de São Paulo com ao menos 1 caso de COVID-19. Destes, a situação de leitos obstétricos particulares foi a seguinte:
  - O 1º quartil dos 383 municípios não possuía leitos obstétricos particulares.
  - O 2º quartil dos 383 municípios possuía até 1 leito obstétrico particular.
  - O 3º quartil dos 383 municípios possuía até 6 leitos obstétricos particulares.
  - A cidade com maior oferta de leitos obstétricos particulares foi São Paulo, com 1045 (o outlier no boxplot, que representa 44% da soma de todos os outros 382 municípios), ou seja, São Paulo manteve o mesmo nº de leitos obstétricos particulares em relação ao mês anterior.
  - Os 383 municípios paulistas com ao menos 1 caso de COVID-19 no mês 5 possuíam ao todo 3390 leitos obstétricos particulares.
  - Os altos valores de amplitude (1045), variância (3002) e coeficiente de variação (619%) demonstram como o nº de leitos obstétricos particulares nos municípios está espalhado e distante da média (8).

	Leitos Obstétricos Não SUS
Contagem	383
Média	8.85117
Desvio Padrão	54.79401
Mínimo	0
25%	0
50%	1
75%	6
Máximo	1045
Soma	3390
Moda	0
Mediana	1
Amplitude	1045
Variância	3002.38355
Coeficiente de Variação	619.05917



# 4. Análise Exploratória de Dados

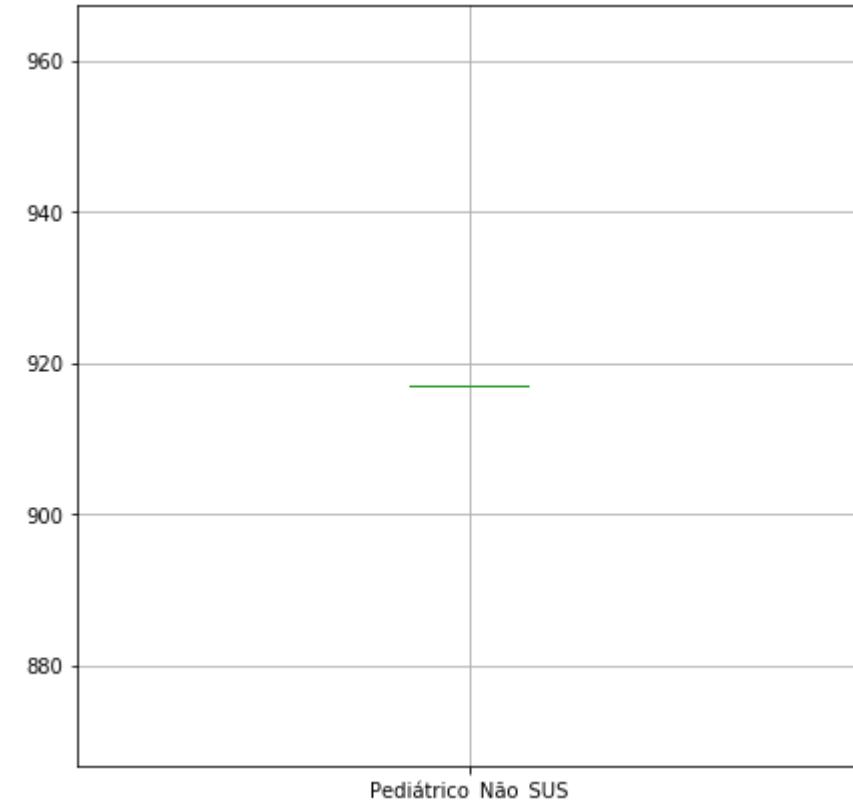
RAIO-X DA BASE | ANÁLISE UNIVARIADA

244

## 'Leitos Pediátricos Não SUS' – mês 2

- Obs.: não faz sentido fazer análise univariada de nº de leitos hospitalares, pois eles são únicos para cada cidade e para cada mês. Portanto, vamos analisar o nº de leitos por mês.
- No mês 2, só havia observações da cidade de São Paulo, que apresentava 917 leitos pediátricos particulares.

	Leitos Pediátricos Não SUS
Contagem	1
Média	917
Desvio Padrão	-
Mínimo	917
25%	917
50%	917
75%	917
Máximo	917
Soma	917
Moda	917
Mediana	917
Amplitude	0
Variância	-
Coeficiente de Variação	-



# 4. Análise Exploratória de Dados

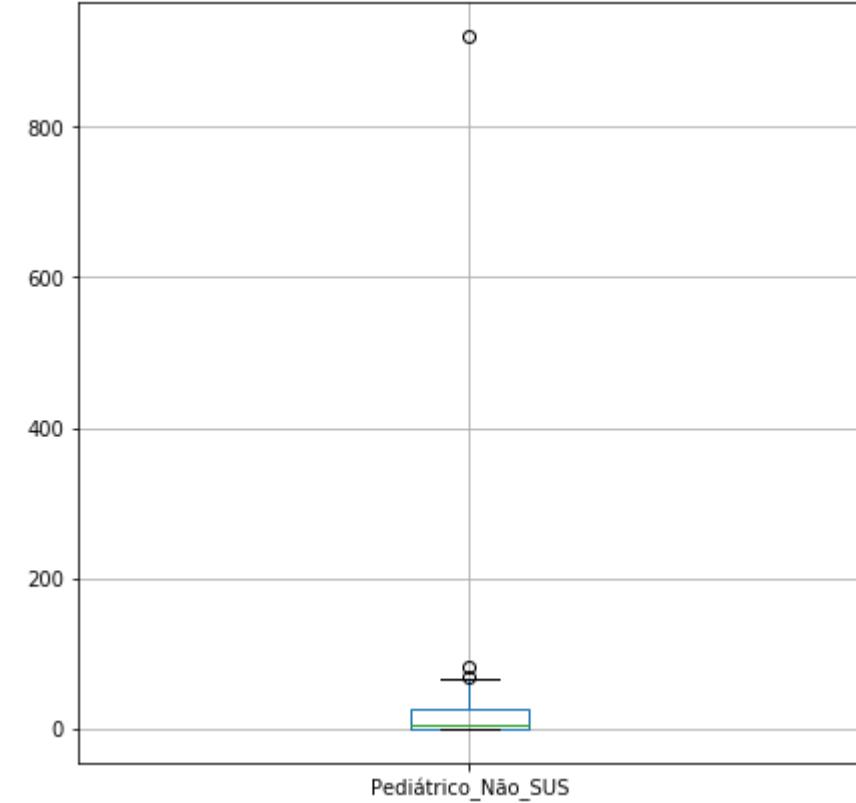
RAIO-X DA BASE | ANÁLISE UNIVARIADA

245

## 'Leitos Pediátricos Não SUS' – mês 3

- No mês 3, havia 67 municípios do estado de São Paulo com ao menos 1 caso de COVID-19. Destes, a situação de leitos pediátricos particulares foi a seguinte:
  - O 1º quartil dos 67 municípios não possuía leitos pediátricos particulares.
  - O 2º quartil dos 67 municípios possuía até 5 leitos pediátricos particulares.
  - O 3º quartil dos 67 municípios possuía até 27 leitos pediátricos particulares.
  - A cidade com maior oferta de leitos pediátricos particulares foi São Paulo, com 920 (o outlier no boxplot, que representa 85% da soma de todos os outros 66 municípios), o que denota um leve aumento em relação ao mês anterior.
  - Os 67 municípios paulistas com ao menos 1 caso de COVID-19 no mês 3 possuíam ao todo 1997 leitos pediátricos particulares.
  - Os altos valores de amplitude (920), variância (12677) e coeficiente de variação (377%) demonstram como o nº de leitos pediátricos particulares nos municípios está espalhado e distante da média (29).

	Leitos Pediátricos Não SUS
Contagem	67
Média	29.80597
Desvio Padrão	112.59235
Mínimo	0
25%	0
50%	5
75%	27.5
Máximo	920
Soma	1997
Moda	0
Mediana	5
Amplitude	920
Variância	12677.03754
Coeficiente de Variação	377.75100



# 4. Análise Exploratória de Dados

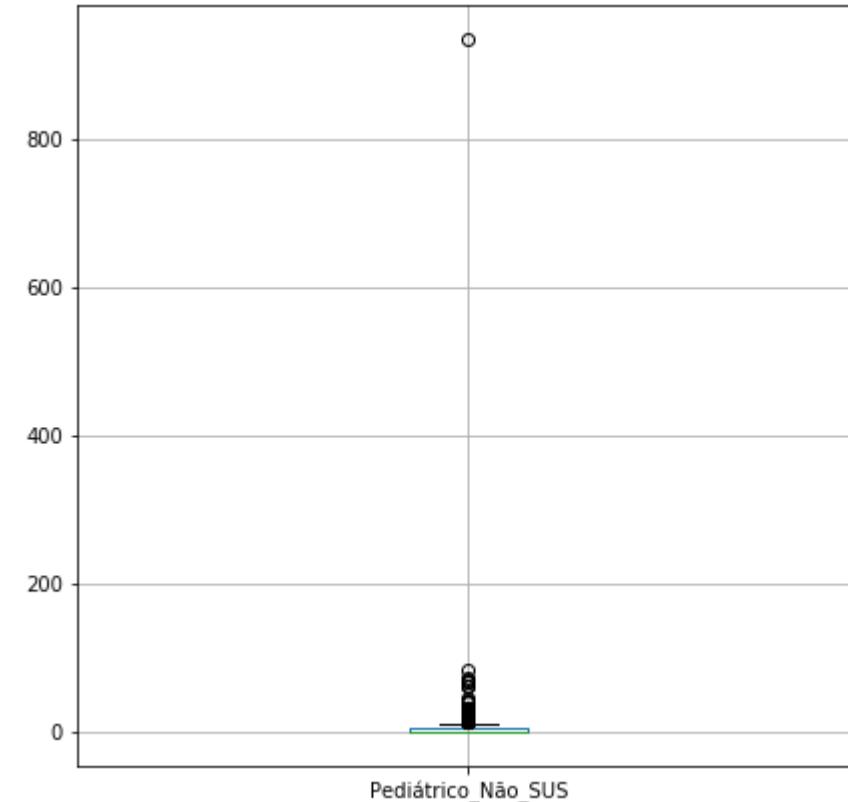
RAIO-X DA BASE | ANÁLISE UNIVARIADA

246

## 'Leitos Pediátricos Não SUS' – mês 4

- No mês 4, havia 321 municípios do estado de São Paulo com ao menos 1 caso de COVID-19. Destes, a situação de leitos pediátricos particulares foi a seguinte:
  - O 1º quartil dos 321 municípios não possuía leitos pediátricos particulares.
  - O 2º quartil dos 321 municípios possuía até 1 leito pediátrico particular.
  - O 3º quartil dos 321 municípios possuía até 5 leitos pediátricos particulares.
  - A cidade com maior oferta de leitos pediátricos particulares foi São Paulo, com 935 (o outlier no boxplot, que representa 50% da soma de todos os outros 320 municípios), o que denota um aumento em relação ao mês anterior.
  - Os 321 municípios paulistas com ao menos 1 caso de COVID-19 no mês 4 possuíam ao todo 2774 leitos pediátricos particulares.
  - Os altos valores de amplitude (935), variância (2854) e coeficiente de variação (618%) demonstram como o nº de leitos pediátricos particulares nos municípios está espalhado e distante da média (8).

Leitos Pediátricos Não SUS	
Contagem	321
Média	8.64174
Desvio Padrão	53.43021
Mínimo	0
25%	0
50%	1
75%	5
Máximo	935
Soma	2774
Moda	0
Mediana	1
Amplitude	935
Variância	2854.78688
Coeficiente de Variação	618.28032



# 4. Análise Exploratória de Dados

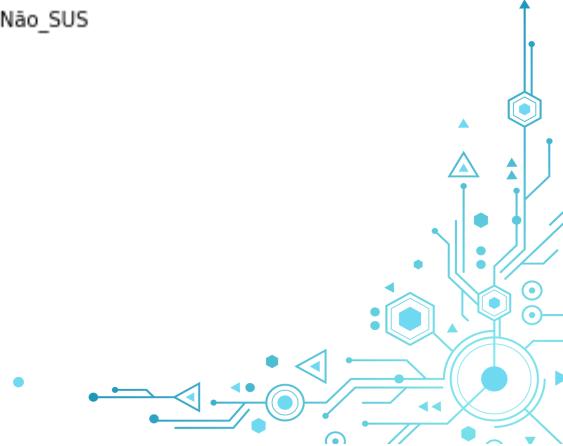
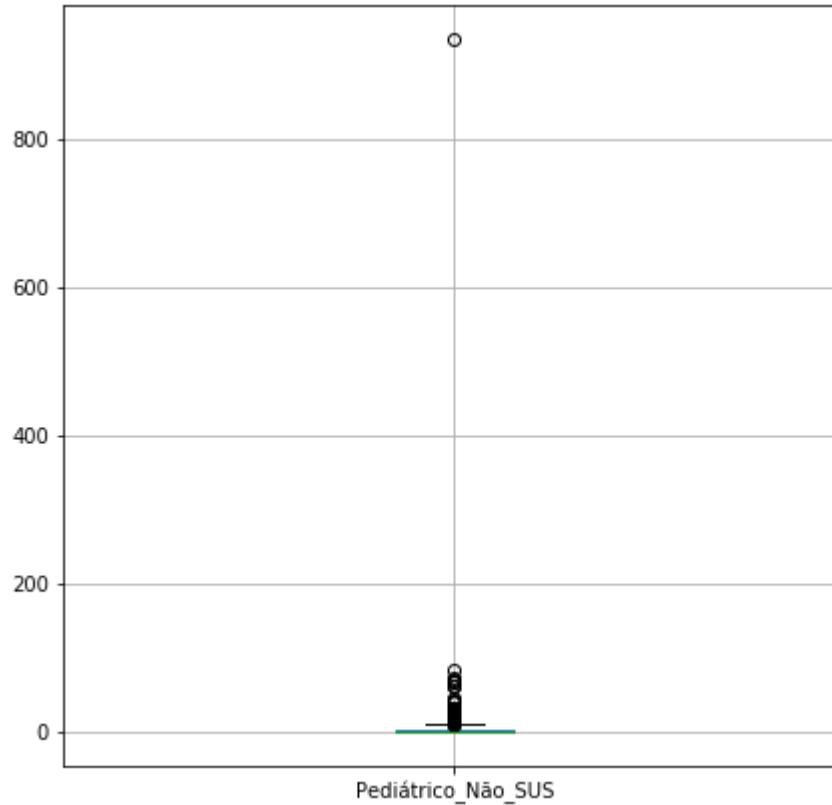
RAIO-X DA BASE | ANÁLISE UNIVARIADA

247

## 'Leitos Pediátricos Não SUS' – mês 5

- No mês 5, havia 383 municípios do estado de São Paulo com ao menos 1 caso de COVID-19. Destes, a situação de leitos pediátricos particulares foi a seguinte:
  - O 1º e o 2º quartil dos 383 municípios não possuía leitos pediátricos particulares.
  - O 3º quartil dos 383 municípios possuía até 4 leitos pediátricos particulares.
  - A cidade com maior oferta de leitos pediátricos particulares foi São Paulo, com 935 (o outlier no boxplot, que representa 49% da soma de todos os outros 382 municípios), ou seja, São Paulo manteve o mesmo nº de leitos pediátricos particulares em relação ao mês anterior.
  - Os 383 municípios paulistas com ao menos 1 caso de COVID-19 no mês 5 possuíam ao todo 2821 leitos pediátricos particulares.
  - Os altos valores de amplitude (935), variância (2400) e coeficiente de variação (665%) demonstram como o nº de leitos pediátricos particulares nos municípios está espalhado e distante da média (7).

Leitos Pediátricos Não SUS	
Contagem	383
Média	7.36554
Desvio Padrão	48.99639
Mínimo	0
25%	0
50%	0
75%	4
Máximo	935
Soma	2821
Moda	0
Mediana	0
Amplitude	935
Variância	2400.64614
Coeficiente de Variação	665.21152



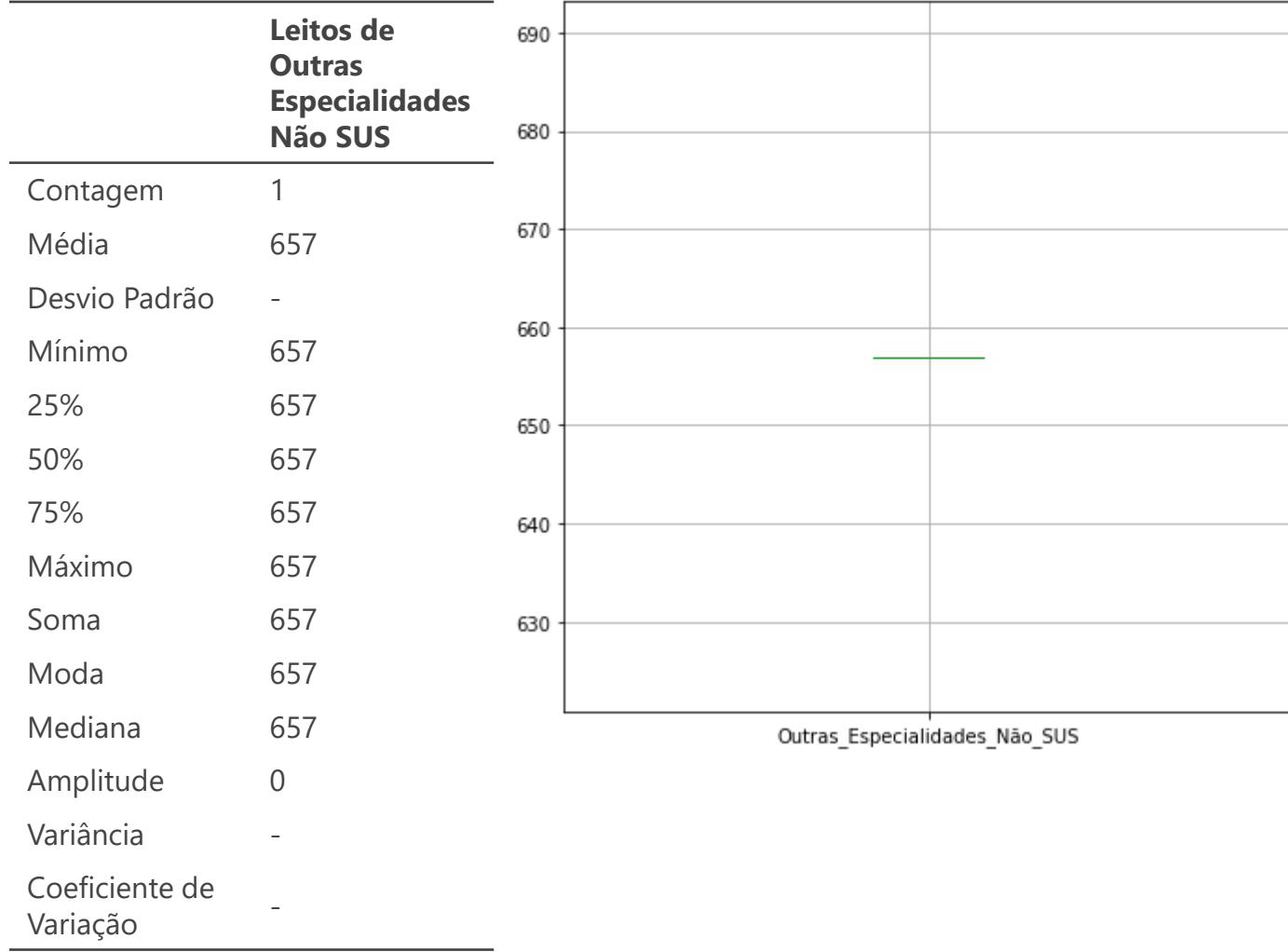
# 4. Análise Exploratória de Dados

RAIO-X DA BASE | ANÁLISE UNIVARIADA

248

## 'Leitos de Outras Especialidades Não SUS' – mês 2

- Obs.: não faz sentido fazer análise univariada de nº de leitos hospitalares, pois eles são únicos para cada cidade e para cada mês. Portanto, vamos analisar o nº de leitos por mês.
- No mês 2, só havia observações da cidade de São Paulo, que apresentava 657 leitos de outras especialidades particulares.



## 4. Análise Exploratória de Dados

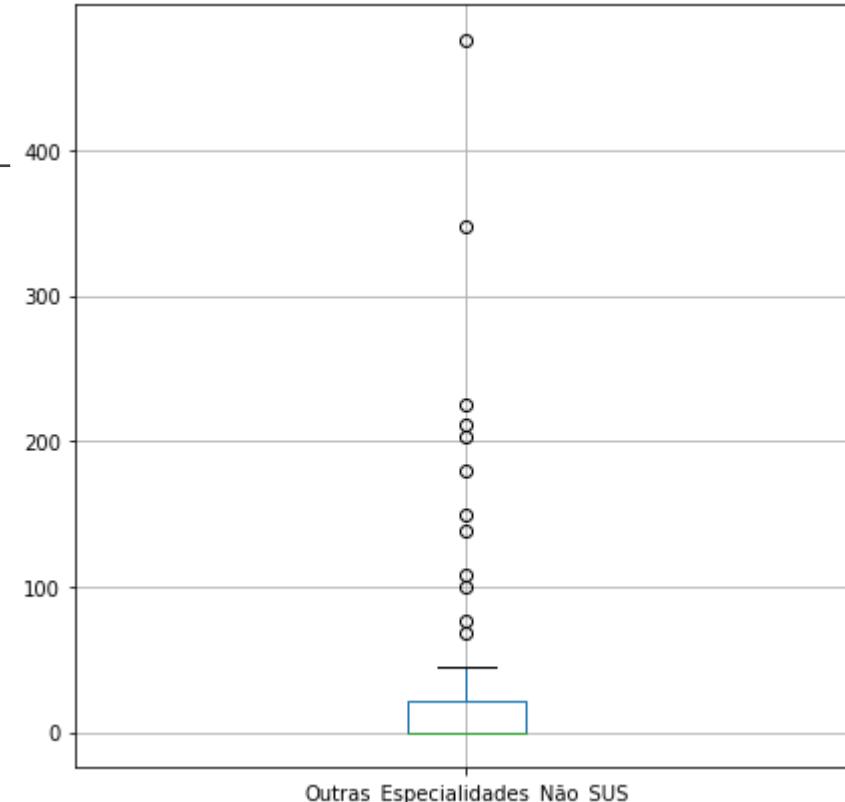
RAIO-X DA BASE | ANÁLISE UNIVARIADA

249

### 'Leitos de Outras Especialidades Não SUS' – mês 3

- No mês 3, havia 67 municípios do estado de São Paulo com ao menos 1 caso de COVID-19. Destes, a situação de leitos de outras especialidades particulares foi a seguinte:
  - O 1º e o 2º quartis dos 67 municípios não possuíam leitos de outras especialidades particulares.
  - O 3º quartil dos 67 municípios possuía até 22 leitos de outras especialidades particulares.
  - A cidade com maior oferta de leitos de outras especialidades particulares foi São Paulo, com 476 (o outlier no boxplot, que representa 22% da soma de todos os outros 66 municípios), o que denota queda em relação ao mês anterior.
  - Os 67 municípios paulistas com ao menos 1 caso de COVID-19 no mês 3 possuíam ao todo 2560 leitos de outras especialidades particulares.
  - Os altos valores de amplitude (476), variância (7563) e coeficiente de variação (227%) demonstram como o nº de leitos de outras especialidades particulares nos municípios está espalhado e distante da média (38).

	Leitos de Outras Especialidades Não SUS
Contagem	67
Média	38.20896
Desvio Padrão	86.96752
Mínimo	0
25%	0
50%	0
75%	22
Máximo	476
Soma	2560
Moda	0
Mediana	0
Amplitude	476
Variância	7563.34962
Coeficiente de Variação	227.61031



# 4. Análise Exploratória de Dados

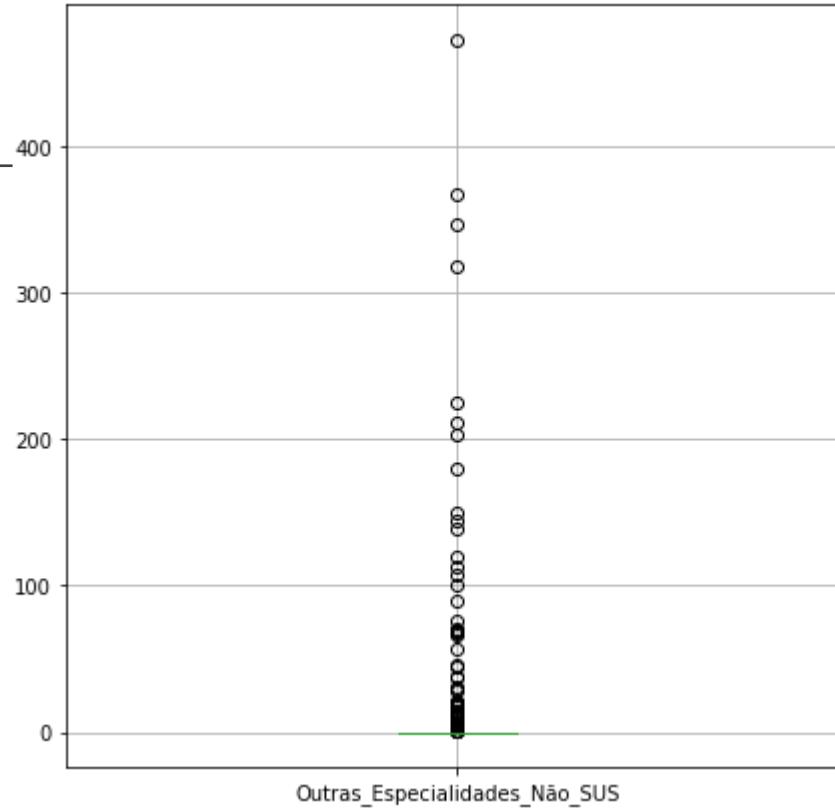
RAIO-X DA BASE | ANÁLISE UNIVARIADA

250

## 'Leitos de Outras Especialidades Não SUS' – mês 4

- No mês 4, havia 321 municípios do estado de São Paulo com ao menos 1 caso de COVID-19. Destes, a situação de leitos de outras especialidades particulares foi a seguinte:
  - O 1º, o 2º e o 3º quartis dos 321 municípios não possuíam leitos de outras especialidades particulares.
  - A cidade com maior oferta de leitos de outras especialidades particulares foi São Paulo, com 473 (o outlier no boxplot, que representa 12% da soma de todos os outros 320 municípios), o que denota queda em relação ao mês anterior.
  - Os 321 municípios paulistas com ao menos 1 caso de COVID-19 no mês 4 possuíam ao todo 4310 leitos de outras especialidades particulares.
  - Os altos valores de amplitude (473), variância (2677) e coeficiente de variação (385%) demonstram como o nº de leitos de outras especialidades particulares nos municípios está espalhado e distante da média (13).

Leitos de Outras Especialidades Não SUS	
Contagem	321
Média	13.42679
Desvio Padrão	51.74784
Mínimo	0
25%	0
50%	0
75%	0
Máximo	473
Soma	4310
Moda	0
Mediana	0
Amplitude	473
Variância	2677.83915
Coeficiente de Variação	385.40736



## 4. Análise Exploratória de Dados

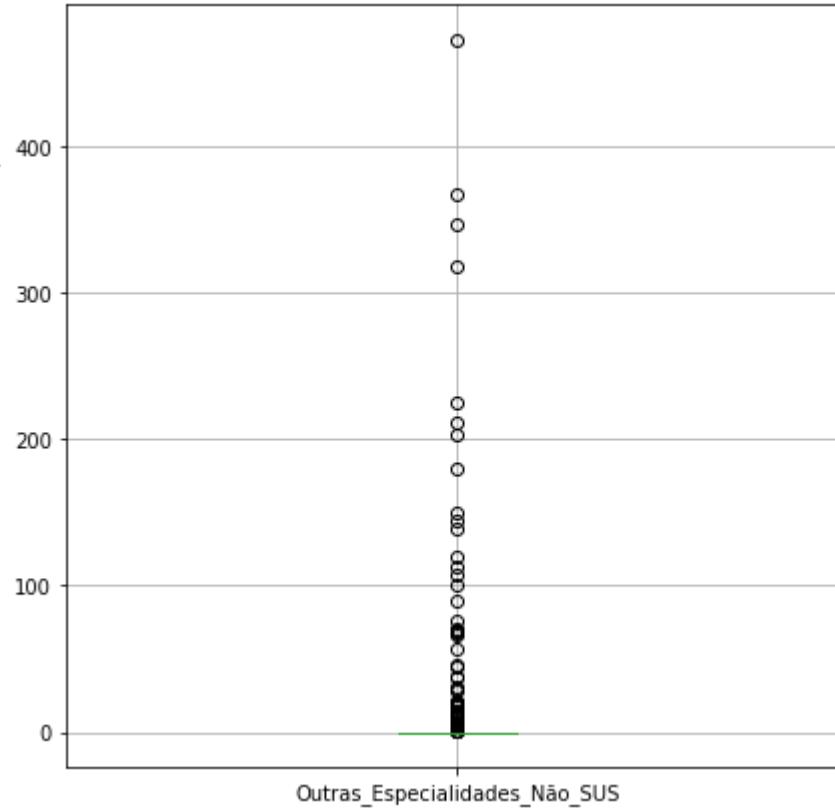
RAIO-X DA BASE | ANÁLISE UNIVARIADA

251

### 'Leitos de Outras Especialidades Não SUS' – mês 5

- No mês 5, havia 383 municípios do estado de São Paulo com ao menos 1 caso de COVID-19. Destes, a situação de leitos de outras especialidades particulares foi a seguinte:
  - O 1º, o 2º e o 3º quartis dos 383 municípios não possuíam leitos de outras especialidades particulares.
  - A cidade com maior oferta de leitos de outras especialidades particulares foi São Paulo, com 473 (o outlier no boxplot, que representa 12% da soma de todos os outros 382 municípios), mantendo o mesmo nº do mês anterior.
  - Os 383 municípios paulistas com ao menos 1 caso de COVID-19 no mês 5 possuíam ao todo 4310 leitos de outras especialidades particulares.
  - Os altos valores de amplitude (473), variância (2267) e coeficiente de variação (423%) demonstram como o nº de leitos de outras especialidades particulares nos municípios está espalhado e distante da média (11).

Leitos de Outras Especialidades Não SUS	
Contagem	383
Média	11.25326
Desvio Padrão	47.62079
Mínimo	0
25%	0
50%	0
75%	0
Máximo	473
Soma	4310
Moda	0
Mediana	0
Amplitude	473
Variância	2267.73935
Coeficiente de Variação	423.17312



# 4. Análise Exploratória de Dados

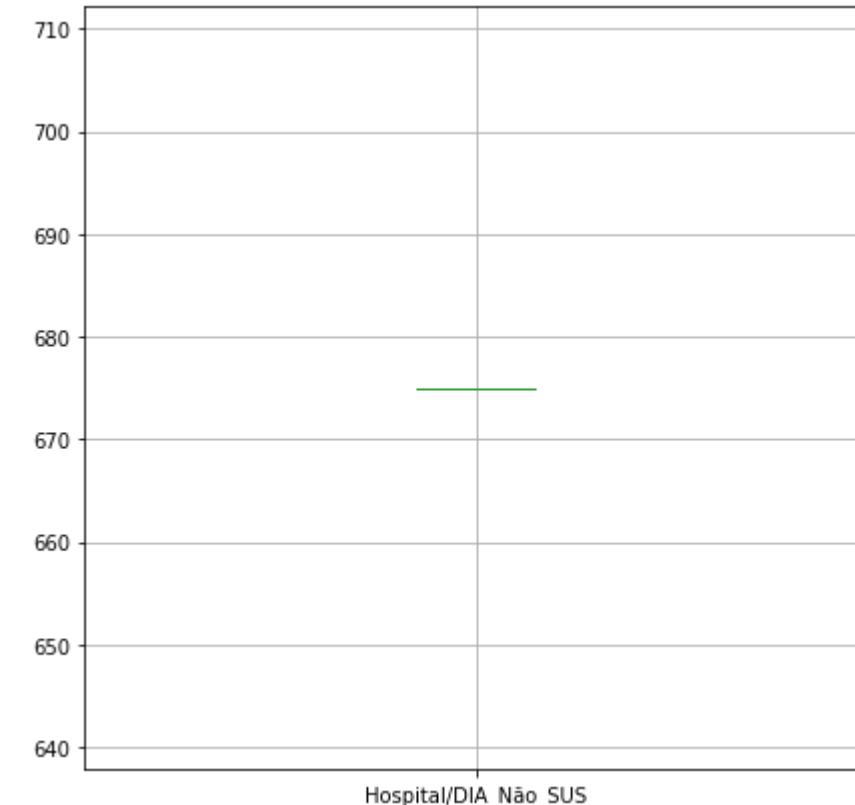
RAIO-X DA BASE | ANÁLISE UNIVARIADA

252

## 'Leitos Hospital/Dia Não SUS' – mês 2

- Obs.: não faz sentido fazer análise univariada de nº de leitos hospitalares, pois eles são únicos para cada cidade e para cada mês. Portanto, vamos analisar o nº de leitos por mês.
- No mês 2, só havia observações da cidade de São Paulo, que apresentava 675 leitos Hospital/DIA particulares.

	Leitos Hospital/Dia Não SUS
Contagem	1
Média	675
Desvio Padrão	-
Mínimo	675
25%	675
50%	675
75%	675
Máximo	675
Soma	675
Moda	675
Mediana	675
Amplitude	0
Variância	-
Coeficiente de Variação	-



## 4. Análise Exploratória de Dados

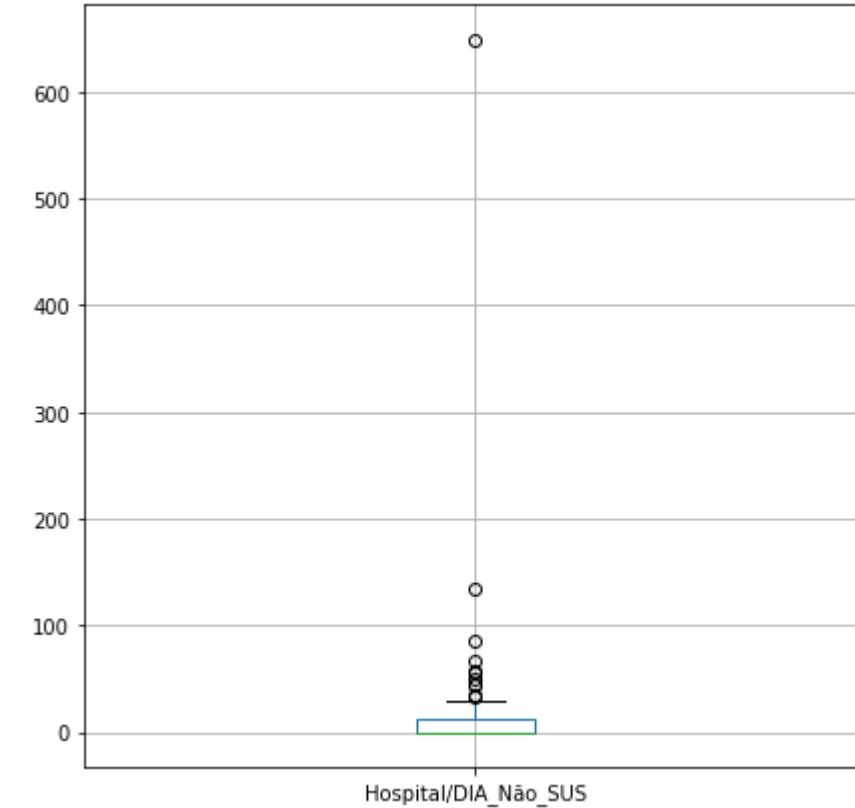
RAIO-X DA BASE | ANÁLISE UNIVARIADA

253

### 'Leitos Hospital/Dia Não SUS' – mês 3

- No mês 3, havia 67 municípios do estado de São Paulo com ao menos 1 caso de COVID-19. Destes, a situação de leitos Hospital/DIA particulares foi a seguinte:
  - O 1º e o 2º quartis dos 67 municípios não possuíam leitos Hospital/DIA particulares.
  - O 3º quartil dos 67 municípios possuía até 12 leitos Hospital/DIA particulares.
  - A cidade com maior oferta de leitos Hospital/DIA particulares foi São Paulo, com 649 (o outlier no boxplot, que representa 84% da soma de todos os outros 66 municípios), o que denota queda em relação ao mês anterior.
  - Os 67 municípios paulistas com ao menos 1 caso de COVID-19 no mês 3 possuíam ao todo 1418 leitos Hospital/DIA particulares.
  - Os altos valores de amplitude (649), variância (6646) e coeficiente de variação (385%) demonstram como o nº de leitos Hospital/DIA particulares nos municípios está espalhado e distante da média (21).

	Leitos Hospital/Dia Não SUS
Contagem	67
Média	21.16418
Desvio Padrão	81.52404
Mínimo	0
25%	0
50%	0
75%	12.5
Máximo	649
Soma	1418
Moda	0
Mediana	0
Amplitude	649
Variância	6646.16961
Coeficiente de Variação	385.19823



# 4. Análise Exploratória de Dados

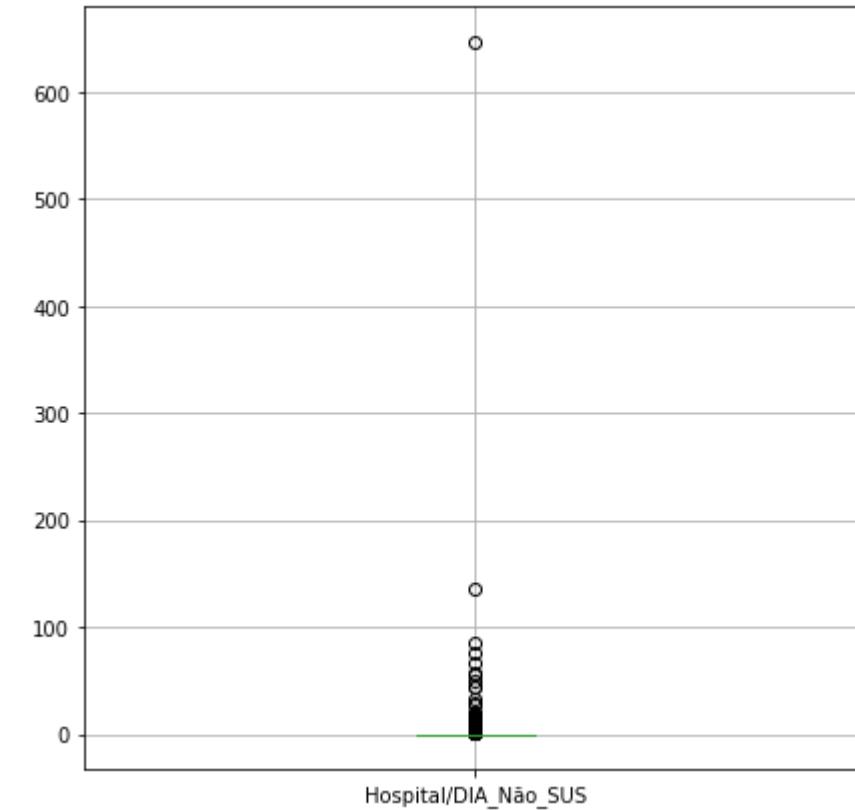
RAIO-X DA BASE | ANÁLISE UNIVARIADA

254

## 'Leitos Hospital/Dia Não SUS' – mês 4

- No mês 4, havia 321 municípios do estado de São Paulo com ao menos 1 caso de COVID-19. Destes, a situação de leitos Hospital/DIA particulares foi a seguinte:
- O 1º, o 2º e o 3º quartis dos 321 municípios não possuíam leitos Hospital/DIA particulares.
- A cidade com maior oferta de leitos Hospital/DIA particulares foi São Paulo, com 647 (o outlier no boxplot, que representa 62% da soma de todos os outros 320 municípios), o que denota queda em relação ao mês anterior.
- Os 321 municípios paulistas com ao menos 1 caso de COVID-19 no mês 4 possuíam ao todo 1680 leitos Hospital/DIA particulares.
- Os altos valores de amplitude (647), variância (1451) e coeficiente de variação (728%) demonstram como o nº de leitos Hospital/DIA particulares nos municípios está espalhado e distante da média (5).

	Leitos Hospital/Dia Não SUS
Contagem	321
Média	5.23364
Desvio Padrão	38.10132
Mínimo	0
25%	0
50%	0
75%	0
Máximo	647
Soma	1680
Moda	0
Mediana	0
Amplitude	647
Variância	1451.71086
Coeficiente de Variação	728.00743



## 4. Análise Exploratória de Dados

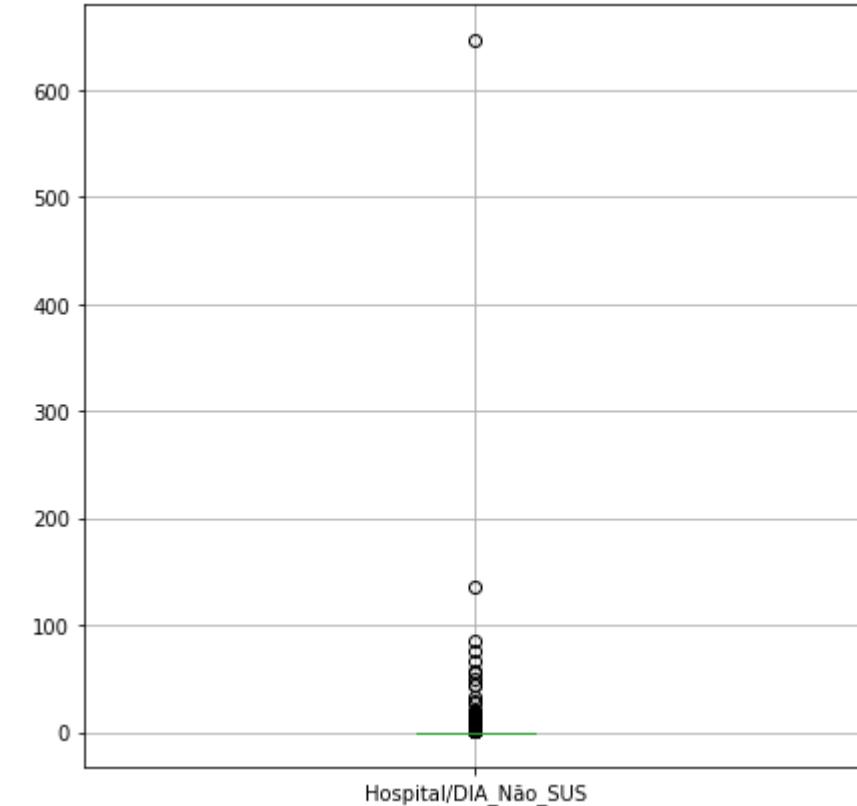
RAIO-X DA BASE | ANÁLISE UNIVARIADA

255

### 'Leitos Hospital/Dia Não SUS' – mês 5

- No mês 5, havia 383 municípios do estado de São Paulo com ao menos 1 caso de COVID-19. Destes, a situação de leitos Hospital/DIA particulares foi a seguinte:
  - O 1º, o 2º e o 3º quartis dos 383 municípios não possuía leitos Hospital/DIA particulares.
  - A cidade com maior oferta de leitos Hospital/DIA particulares foi São Paulo, com 647 (o outlier no boxplot, que representa 62% da soma de todos os outros 382 municípios), ou seja, São Paulo manteve o mesmo nº de leitos Hospital/DIA particulares em relação ao mês anterior.
  - Os 383 municípios paulistas com ao menos 1 caso de COVID-19 no mês 5 possuíam ao todo 1682 leitos Hospital/DIA particulares.
  - Os altos valores de amplitude (647), variância (1219) e coeficiente de variação (795%) demonstram como o nº de leitos Hospital/DIA particulares nos municípios está espalhado e distante da média (4).

	Leitos Hospital/Dia Não SUS
Contagem	383
Média	4.39164
Desvio Padrão	34.92540
Mínimo	0
25%	0
50%	0
75%	0
Máximo	647
Soma	1682
Moda	0
Mediana	0
Amplitude	647
Variância	1219.78339
Coeficiente de Variação	795.26916



# 4. Análise Exploratória de Dados

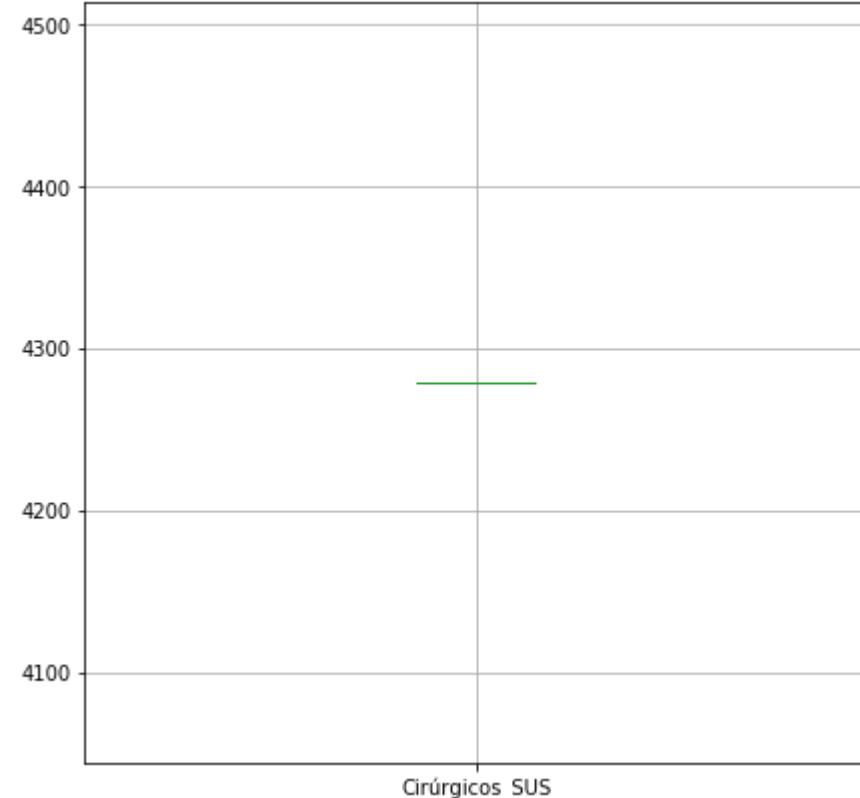
RAIO-X DA BASE | ANÁLISE UNIVARIADA

256

## 'Leitos Cirúrgicos SUS' – mês 2

- Obs.: não faz sentido fazer análise univariada de nº de leitos hospitalares, pois eles são únicos para cada cidade e para cada mês. Portanto, vamos analisar o nº de leitos por mês.
- No mês 2, só havia observações da cidade de São Paulo, que apresentava 4279 leitos cirúrgicos no SUS.

	Leitos Cirúrgicos SUS
Contagem	1
Média	4279
Desvio Padrão	-
Mínimo	4279
25%	4279
50%	4279
75%	4279
Máximo	4279
Soma	4279
Moda	4279
Mediana	4279
Amplitude	0
Variância	-
Coeficiente de Variação	-



# 4. Análise Exploratória de Dados

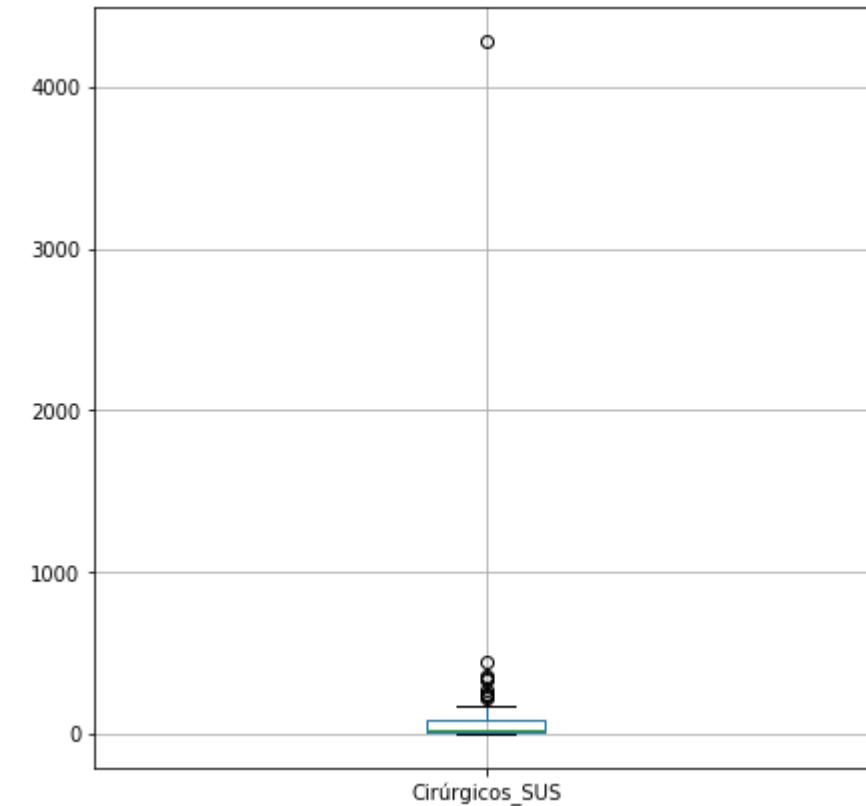
RAIO-X DA BASE | ANÁLISE UNIVARIADA

257

## 'Leitos Cirúrgicos SUS' – mês 3

- No mês 3, havia 67 municípios do estado de São Paulo com ao menos 1 caso de COVID-19. Destes, a situação de leitos cirúrgicos do SUS foi a seguinte:
  - O 1º quartil dos 67 municípios possuía até 8 leitos cirúrgicos do SUS.
  - O 2º quartil dos 67 municípios possuía até 27 leitos cirúrgicos do SUS.
  - O 3º quartil dos 67 municípios possuía até 84 leitos cirúrgicos do SUS.
  - A cidade com maior oferta de leitos cirúrgicos do SUS foi São Paulo, com 4279 (o outlier no boxplot, que representa 87% da soma de todos os outros 66 municípios), mesmo número do mês anterior.
  - Os 67 municípios paulistas com ao menos 1 caso de COVID-19 no mês 3 possuíam ao todo 9180 leitos cirúrgicos do SUS.
  - Os altos valores de amplitude (4279), variância (274548) e coeficiente de variação (382%) demonstram como o nº de leitos cirúrgicos do SUS nos municípios está espalhado e distante da média (137).

	Leitos Cirúrgicos SUS
Contagem	67
Média	137.01493
Desvio Padrão	523.97385
Mínimo	0
25%	8.5
50%	27
75%	84
Máximo	4279
Soma	9180
Moda	0
Mediana	27
Amplitude	4279
Variância	274548.59068
Coeficiente de Variação	382.42100



# 4. Análise Exploratória de Dados

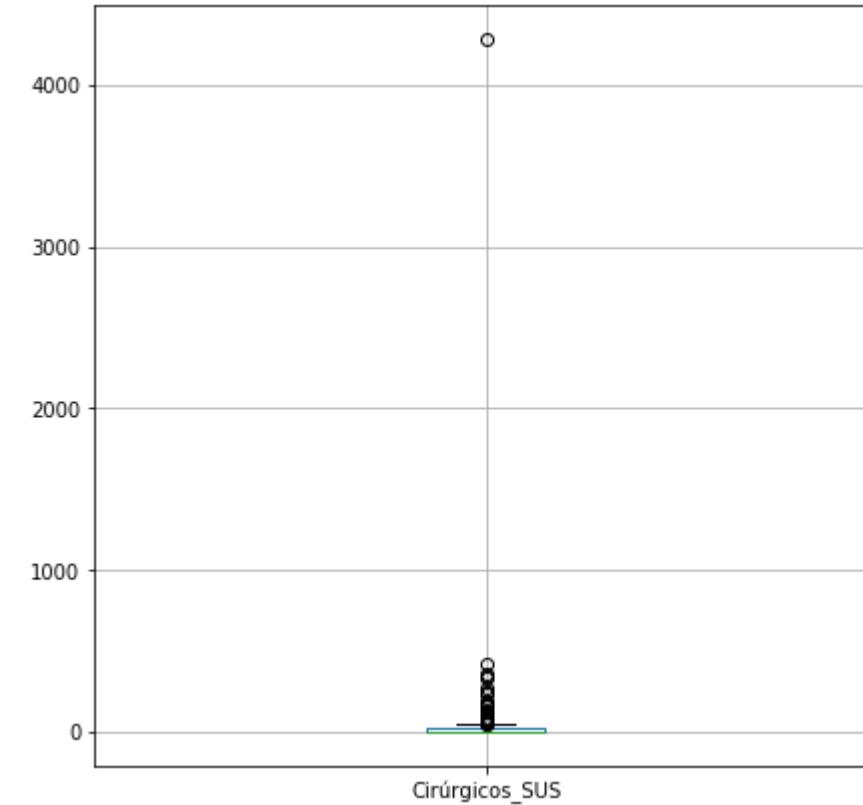
RAIO-X DA BASE | ANÁLISE UNIVARIADA

258

## 'Leitos Cirúrgicos SUS' – mês 4

- No mês 4, havia 321 municípios do estado de São Paulo com ao menos 1 caso de COVID-19. Destes, a situação de leitos cirúrgicos do SUS foi a seguinte:
  - O 1º quartil dos 321 municípios não possuía leitos cirúrgicos do SUS.
  - O 2º quartil dos 321 municípios possuía até 6 leitos cirúrgicos do SUS.
  - O 3º quartil dos 321 municípios possuía até 20 leitos cirúrgicos do SUS.
  - A cidade com maior oferta de leitos cirúrgicos do SUS foi São Paulo, com 4279 (o outlier no boxplot, que representa 52% da soma de todos os outros 320 municípios), mesmo nº do mês anterior.
  - Os 321 municípios paulistas com ao menos 1 caso de COVID-19 no mês 4 possuíam ao todo 12458 leitos cirúrgicos do SUS.
  - Os altos valores de amplitude (4279), variância (59621) e coeficiente de variação (629%) demonstram como o nº de leitos cirúrgicos do SUS nos municípios está espalhado e distante da média (38).

	Leitos Cirúrgicos SUS
Contagem	321
Média	38.80997
Desvio Padrão	244.17556
Mínimo	0
25%	0
50%	6
75%	20
Máximo	4279
Soma	12458
Moda	0
Mediana	6
Amplitude	4279
Variância	59621.70440
Coeficiente de Variação	629.15681



# 4. Análise Exploratória de Dados

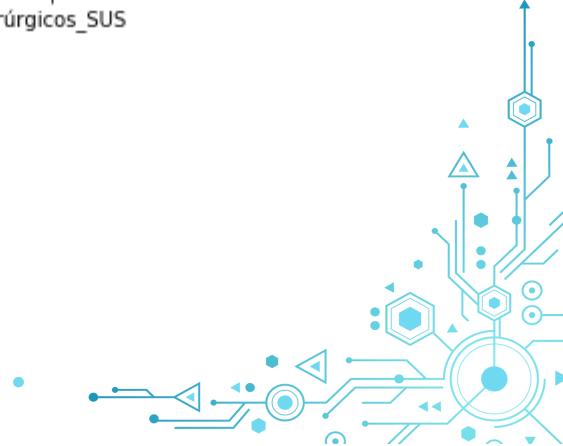
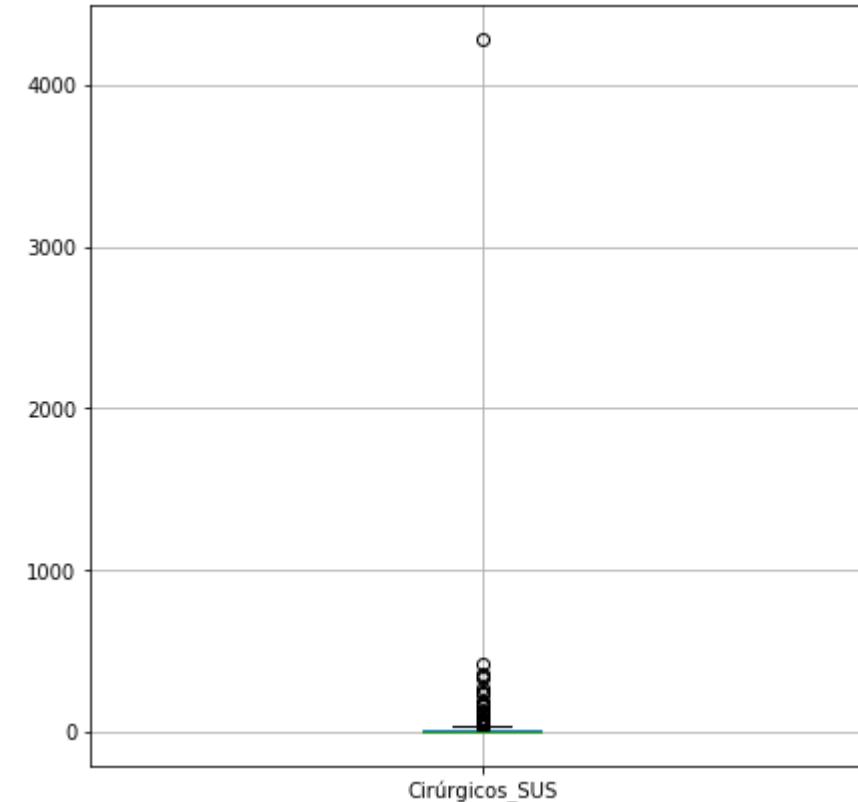
RAIO-X DA BASE | ANÁLISE UNIVARIADA

259

## 'Leitos Cirúrgicos SUS' – mês 5

- No mês 5, havia 383 municípios do estado de São Paulo com ao menos 1 caso de COVID-19. Destes, a situação de leitos cirúrgicos do SUS foi a seguinte:
  - O 1º quartil dos 383 municípios não possuía leitos cirúrgicos do SUS.
  - O 2º quartil dos 383 municípios possuía até 5 leitos cirúrgicos do SUS.
  - O 3º quartil dos 383 municípios possuía até 16 leitos cirúrgicos do SUS.
  - A cidade com maior oferta de leitos cirúrgicos do SUS foi São Paulo, com 4279 (o outlier no boxplot, que representa 51% da soma de todos os outros 382 municípios), ou seja, São Paulo manteve o mesmo nº de leitos cirúrgicos do SUS em relação ao mês anterior.
  - Os 383 municípios paulistas com ao menos 1 caso de COVID-19 no mês 5 possuíam ao todo 12599 leitos cirúrgicos do SUS.
  - Os altos valores de amplitude (4279), variância (50128) e coeficiente de variação (680%) demonstram como o nº de leitos cirúrgicos do SUS nos municípios está espalhado e distante da média (32).

Leitos Cirúrgicos SUS	
Contagem	383
Média	32.89556
Desvio Padrão	223.89308
Mínimo	0
25%	0
50%	5
75%	16
Máximo	4279
Soma	12599
Moda	0
Mediana	5
Amplitude	4279
Variância	50128.10948
Coeficiente de Variação	680.61789



# 4. Análise Exploratória de Dados

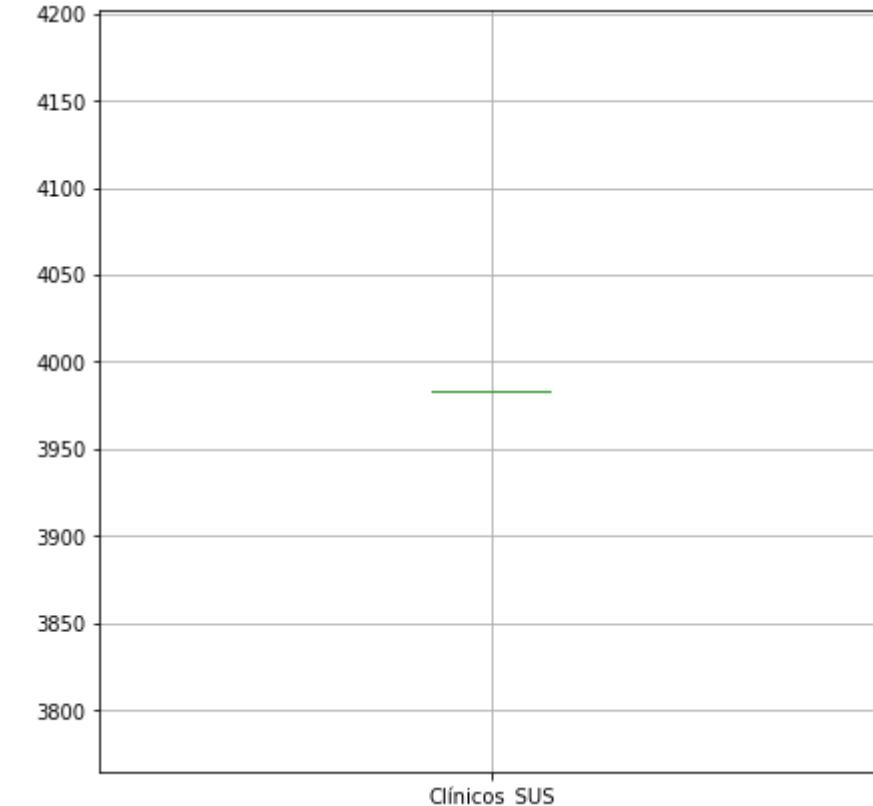
RAIO-X DA BASE | ANÁLISE UNIVARIADA

260

## 'Leitos Clínicos SUS' – mês 2

- Obs.: não faz sentido fazer análise univariada de nº de leitos hospitalares, pois eles são únicos para cada cidade e para cada mês. Portanto, vamos analisar o nº de leitos por mês.
- No mês 2, só havia observações da cidade de São Paulo, que apresentava 3983 leitos clínicos do SUS.

Leitos Clínicos SUS	
Contagem	1
Média	3983
Desvio Padrão	-
Mínimo	3983
25%	3983
50%	3983
75%	3983
Máximo	3983
Soma	3983
Moda	3983
Mediana	3983
Amplitude	0
Variância	-
Coeficiente de Variação	-



# 4. Análise Exploratória de Dados

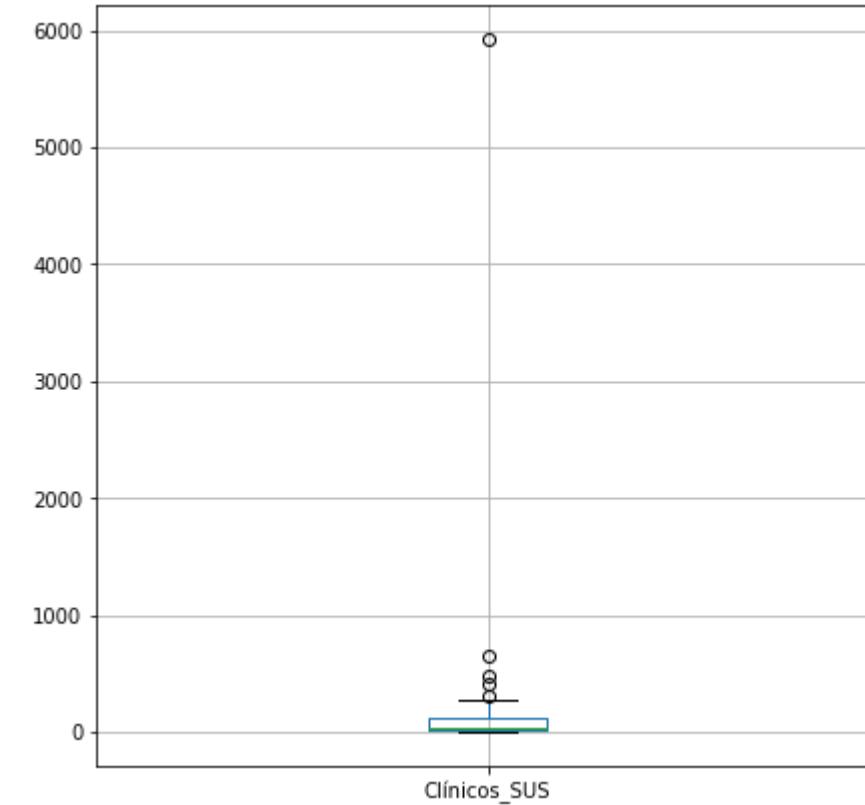
RAIO-X DA BASE | ANÁLISE UNIVARIADA

261

## 'Leitos Clínicos SUS' – mês 3

- No mês 3, havia 67 municípios do estado de São Paulo com ao menos 1 caso de COVID-19. Destes, a situação de leitos clínicos do SUS foi a seguinte:
  - O 1º quartil dos 67 municípios possuía até 21 leitos clínicos do SUS.
  - O 2º quartil dos 67 municípios possuía até 39 leitos clínicos do SUS.
  - O 3º quartil dos 67 municípios possuía até 127 leitos clínicos do SUS.
  - A cidade com maior oferta de leitos clínicos do SUS foi São Paulo, com 5921 (o outlier no boxplot, que representa 96% da soma de todos os outros 66 municípios), o que denota aumento em relação ao mês anterior.
  - Os 67 municípios paulistas com ao menos 1 caso de COVID-19 no mês 3 possuíam ao todo 12062 leitos clínicos do SUS.
  - Os altos valores de amplitude (5921), variância (521909) e coeficiente de variação (401%) demonstram como o nº de leitos clínicos do SUS nos municípios está espalhado e distante da média (180).

Leitos Clínicos SUS	
Contagem	67
Média	180.02985
Desvio Padrão	722.43332
Mínimo	0
25%	21
50%	39
75%	127.5
Máximo	5921
Soma	12062
Moda	21
Mediana	39
Amplitude	5921
Variância	521909.90819
Coeficiente de Variação	401.28530



# 4. Análise Exploratória de Dados

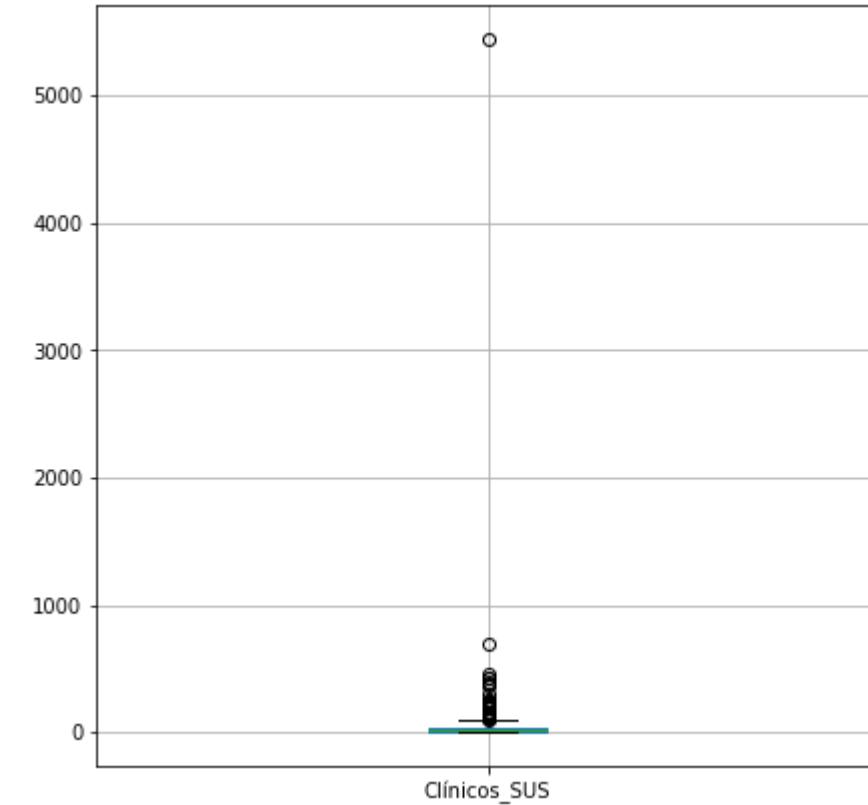
RAIO-X DA BASE | ANÁLISE UNIVARIADA

262

## 'Leitos Clínicos SUS' – mês 4

- No mês 4, havia 321 municípios do estado de São Paulo com ao menos 1 caso de COVID-19. Destes, a situação de leitos clínicos do SUS foi a seguinte:
  - O 1º quartil dos 321 municípios possuía até 5 leitos clínicos do SUS.
  - O 2º quartil dos 321 municípios possuía até 18 leitos clínicos do SUS.
  - O 3º quartil dos 321 municípios possuía até 40 leitos clínicos do SUS.
  - A cidade com maior oferta de leitos clínicos do SUS foi São Paulo, com 5440 (o outlier no boxplot, que representa 39% da soma de todos os outros 320 municípios), o que denota queda em relação ao mês anterior.
  - Os 321 municípios paulistas com ao menos 1 caso de COVID-19 no mês 4 possuíam ao todo 19076 leitos clínicos do SUS.
  - Os altos valores de amplitude (5440), variância (97179) e coeficiente de variação (524%) demonstram como o nº de leitos clínicos do SUS nos municípios está espalhado e distante da média (59).

Leitos Clínicos SUS	
Contagem	321
Média	59.42679
Desvio Padrão	311.73637
Mínimo	0
25%	5
50%	18
75%	40
Máximo	5440
Soma	19076
Moda	0
Mediana	18
Amplitude	5440
Variância	97179.56415
Coeficiente de Variação	524.57210



# 4. Análise Exploratória de Dados

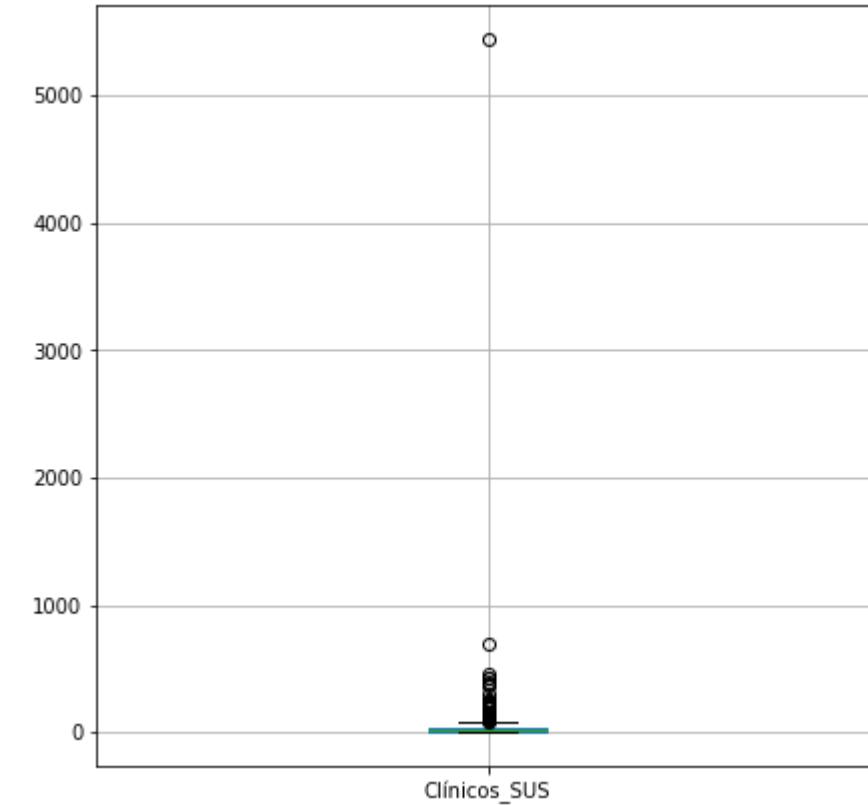
RAIO-X DA BASE | ANÁLISE UNIVARIADA

263

## 'Leitos Clínicos SUS' – mês 5

- No mês 5, havia 383 municípios do estado de São Paulo com ao menos 1 caso de COVID-19. Destes, a situação de leitos clínicos do SUS foi a seguinte:
  - O 1º quartil dos 383 municípios não possuía leitos clínicos do SUS.
  - O 2º quartil dos 383 municípios possuía até 15 leitos clínicos do SUS.
  - O 3º quartil dos 383 municípios possuía até 33 leitos clínicos do SUS.
  - A cidade com maior oferta de leitos clínicos do SUS foi São Paulo, com 5440 (o outlier no boxplot, que representa 38% da soma de todos os outros 382 municípios), ou seja, São Paulo manteve o mesmo nº de leitos clínicos do SUS em relação ao mês anterior.
  - Os 383 municípios paulistas com ao menos 1 caso de COVID-19 no mês 5 possuíam ao todo 19516 leitos clínicos do SUS.
  - Os altos valores de amplitude (5440), variância (81790) e coeficiente de variação (561%) demonstram como o nº de leitos clínicos do SUS nos municípios está espalhado e distante da média (50).

Leitos Clínicos SUS	
Contagem	383
Média	50.95561
Desvio Padrão	285.99110
Mínimo	0
25%	0
50%	15
75%	33
Máximo	5440
Soma	19516
Moda	0
Mediana	15
Amplitude	5440
Variância	81790.91164
Coeficiente de Variação	561.25534



# 4. Análise Exploratória de Dados

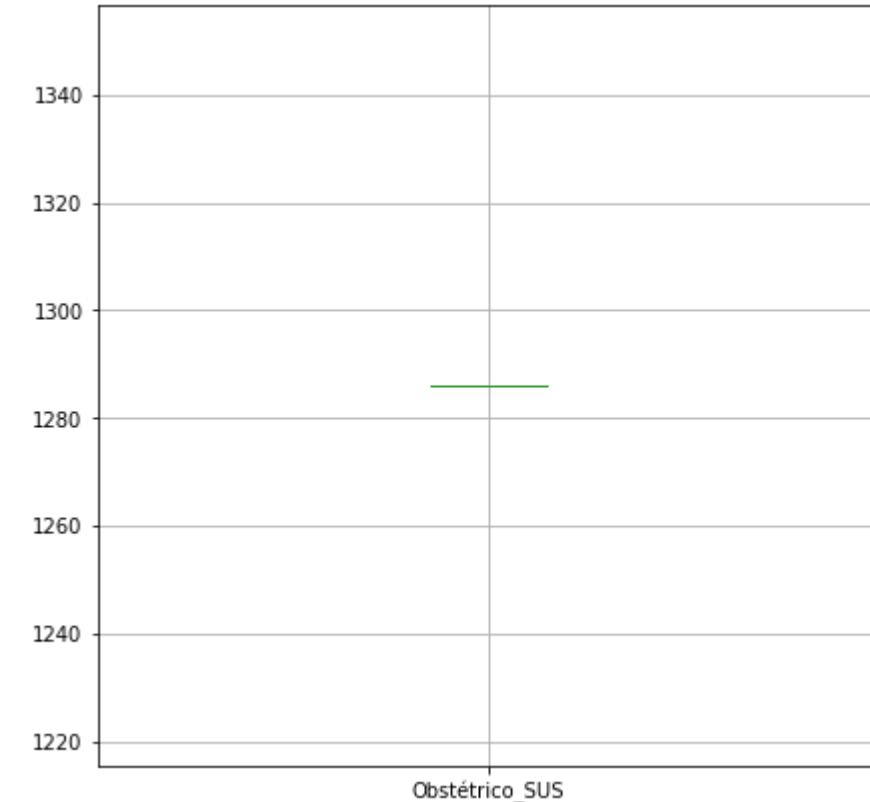
RAIO-X DA BASE | ANÁLISE UNIVARIADA

264

## 'Leitos Obstétricos SUS' – mês 2

- Obs.: não faz sentido fazer análise univariada de nº de leitos hospitalares, pois eles são únicos para cada cidade e para cada mês. Portanto, vamos analisar o nº de leitos por mês.
- No mês 2, só havia observações da cidade de São Paulo, que apresentava 1286 leitos obstétricos do SUS.

	Leitos Obstétricos SUS
Contagem	1
Média	1286
Desvio Padrão	-
Mínimo	1286
25%	1286
50%	1286
75%	1286
Máximo	1286
Soma	1286
Moda	1286
Mediana	1286
Amplitude	0
Variância	-
Coeficiente de Variação	-



# 4. Análise Exploratória de Dados

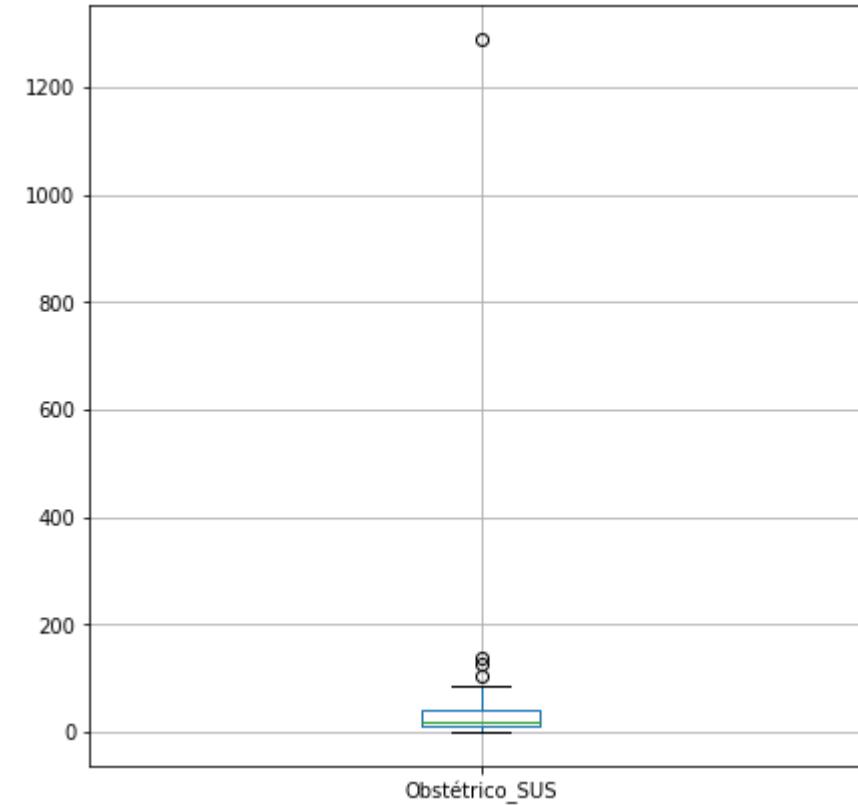
RAIO-X DA BASE | ANÁLISE UNIVARIADA

265

## 'Leitos Obstétricos SUS' – mês 3

- No mês 3, havia 67 municípios do estado de São Paulo com ao menos 1 caso de COVID-19. Destes, a situação de leitos obstétricos do SUS foi a seguinte:
  - O 1º quartil dos 67 municípios possuía até 10 leitos obstétricos do SUS.
  - O 2º quartil dos 67 municípios possuía até 20 leitos obstétricos do SUS.
  - O 3º quartil dos 67 municípios possuía até 41 leitos obstétricos do SUS.
  - A cidade com maior oferta de leitos obstétricos do SUS foi São Paulo, com 1288 (o outlier no boxplot, que representa 63% da soma de todos os outros 66 municípios), o que denota aumento em relação ao mês anterior.
  - Os 67 municípios paulistas com ao menos 1 caso de COVID-19 no mês 3 possuíam ao todo 3304 leitos obstétricos do SUS.
  - Os altos valores de amplitude (1288), variância (24507) e coeficiente de variação (317%) demonstram como o nº de leitos obstétricos do SUS nos municípios está espalhado e distante da média (49).

Leitos Obstétricos SUS	
Contagem	67
Média	49.31343
Desvio Padrão	156.54830
Mínimo	0
25%	10
50%	20
75%	41.5
Máximo	1288
Soma	3304
Moda	0
Mediana	20
Amplitude	1288
Variância	24507.36997
Coeficiente de Variação	317.45569



# 4. Análise Exploratória de Dados

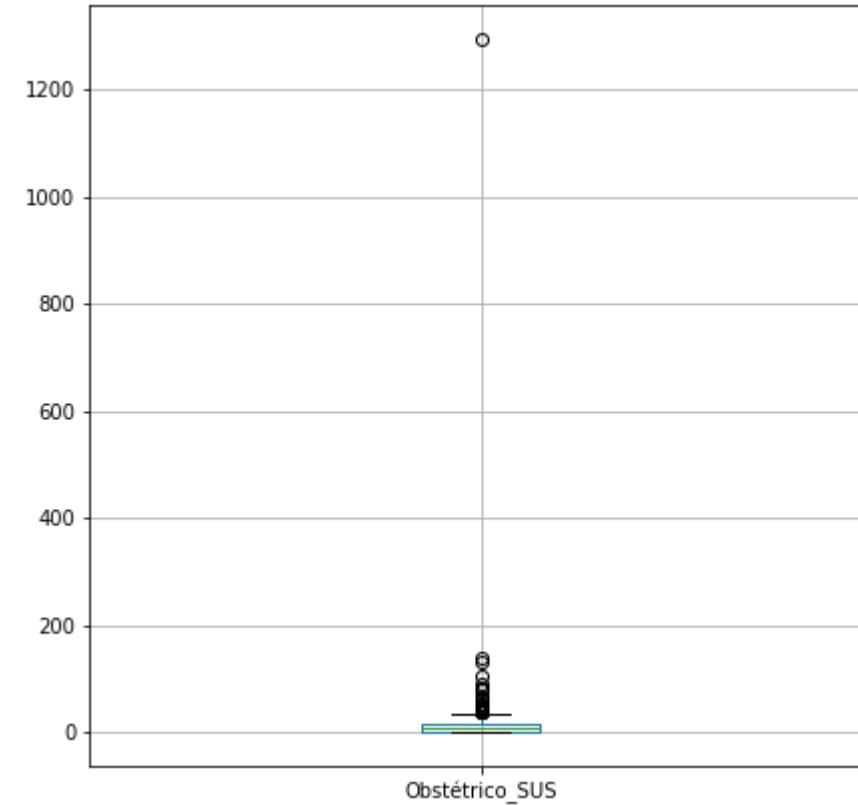
RAIO-X DA BASE | ANÁLISE UNIVARIADA

266

## 'Leitos Obstétricos SUS' – mês 4

- No mês 4, havia 321 municípios do estado de São Paulo com ao menos 1 caso de COVID-19. Destes, a situação de leitos obstétricos do SUS foi a seguinte:
  - O 1º quartil dos 321 municípios não possuía leitos obstétricos do SUS.
  - O 2º quartil dos 321 municípios possuía até 7 leitos obstétricos do SUS.
  - O 3º quartil dos 321 municípios possuía até 14 leitos obstétricos do SUS.
  - A cidade com maior oferta de leitos obstétricos do SUS foi São Paulo, com 1293 (o outlier no boxplot, que representa 1/3 da soma de todos os outros 320 municípios), o que denota aumento em relação ao mês anterior.
  - Os 321 municípios paulistas com ao menos 1 caso de COVID-19 no mês 4 possuíam ao todo 5182 leitos obstétricos do SUS.
  - Os altos valores de amplitude (1293), variância (5458) e coeficiente de variação (457%) demonstram como o nº de leitos obstétricos do SUS nos municípios está espalhado e distante da média (16).

Leitos Obstétricos SUS	
Contagem	321
Média	16.14330
Desvio Padrão	73.88262
Mínimo	0
25%	0
50%	7
75%	14
Máximo	1293
Soma	5182
Moda	0
Mediana	7
Amplitude	1293
Variância	5458.64190
Coeficiente de Variação	457.66734



# 4. Análise Exploratória de Dados

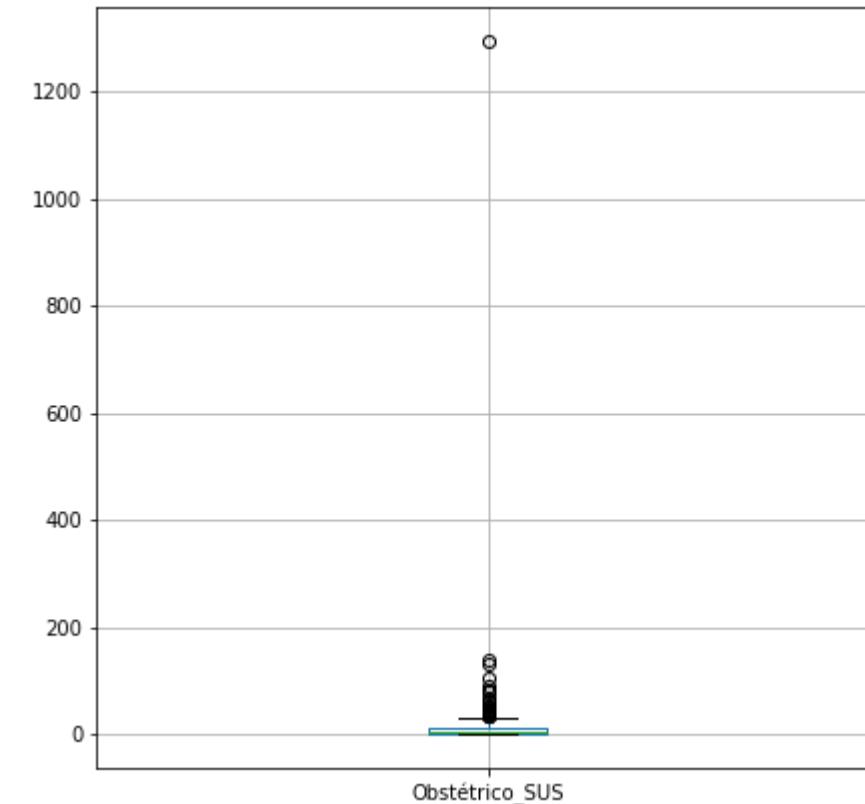
RAIO-X DA BASE | ANÁLISE UNIVARIADA

267

## 'Leitos Obstétricos SUS' – mês 5

- No mês 5, havia 383 municípios do estado de São Paulo com ao menos 1 caso de COVID-19. Destes, a situação de leitos obstétricos do SUS foi a seguinte:
  - O 1º quartil dos 383 municípios não possuía leitos obstétricos do SUS.
  - O 2º quartil dos 383 municípios possuía até 4 leitos obstétricos do SUS.
  - O 3º quartil dos 383 municípios possuía até 13 leitos obstétricos do SUS.
  - A cidade com maior oferta de leitos obstétricos do SUS foi São Paulo, com 1293 (o outlier no boxplot, que representa 32% da soma de todos os outros 382 municípios), ou seja, São Paulo manteve o mesmo nº de leitos obstétricos do SUS em relação ao mês anterior.
  - Os 383 municípios paulistas com ao menos 1 caso de COVID-19 no mês 5 possuíam ao todo 5300 leitos obstétricos do SUS.
  - Os altos valores de amplitude (1293), variância (4601) e coeficiente de variação (490%) demonstram como o nº de leitos obstétricos do SUS nos municípios está espalhado e distante da média (13).

Leitos Obstétricos SUS	
Contagem	383
Média	13.83812
Desvio Padrão	67.83619
Mínimo	0
25%	0
50%	4
75%	13
Máximo	1293
Soma	5300
Moda	0
Mediana	4
Amplitude	1293
Variância	4601.74860
Coeficiente de Variação	490.21246



# 4. Análise Exploratória de Dados

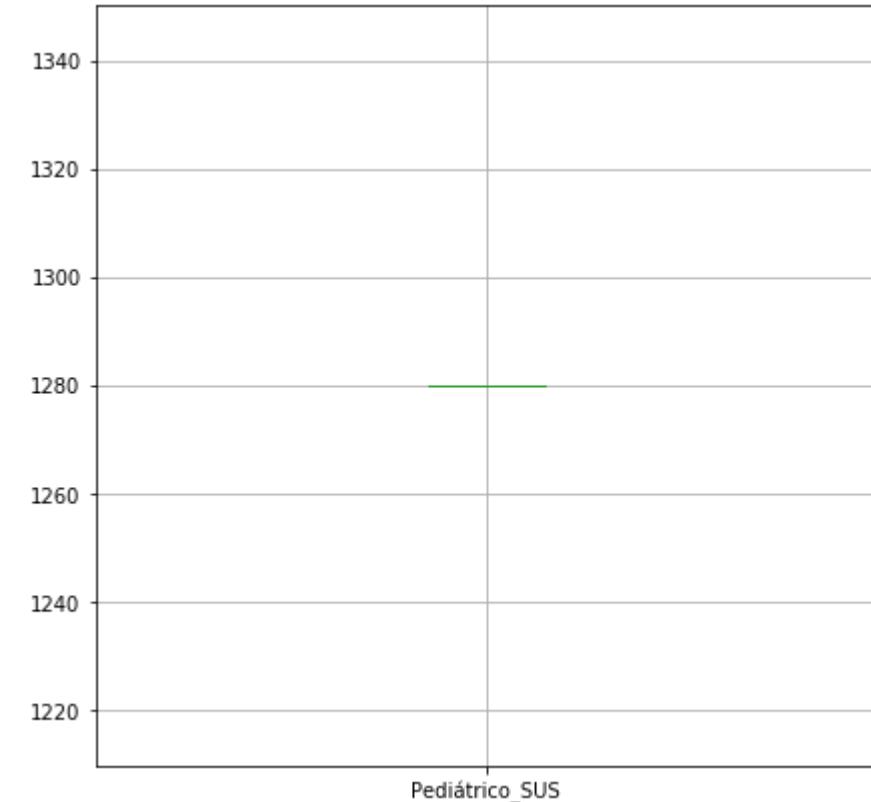
RAIO-X DA BASE | ANÁLISE UNIVARIADA

268

## 'Leitos Pediátricos SUS' – mês 2

- Obs.: não faz sentido fazer análise univariada de nº de leitos hospitalares, pois eles são únicos para cada cidade e para cada mês. Portanto, vamos analisar o nº de leitos por mês.
- No mês 2, só havia observações da cidade de São Paulo, que apresentava 1280 leitos pediátricos do SUS.

	Leitos Pediátricos SUS
Contagem	1
Média	1280
Desvio Padrão	-
Mínimo	1280
25%	1280
50%	1280
75%	1280
Máximo	1280
Soma	1280
Moda	1280
Mediana	1280
Amplitude	0
Variância	-
Coeficiente de Variação	-



# 4. Análise Exploratória de Dados

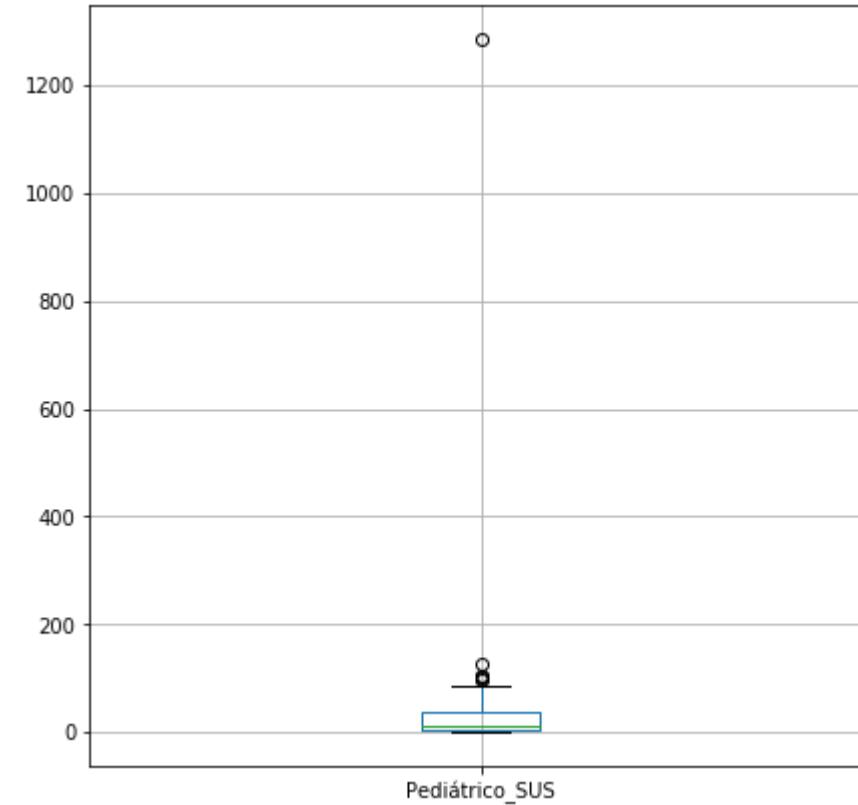
RAIO-X DA BASE | ANÁLISE UNIVARIADA

269

## 'Leitos Pediátricos SUS' – mês 3

- No mês 3, havia 67 municípios do estado de São Paulo com ao menos 1 caso de COVID-19. Destes, a situação de leitos pediátricos do SUS foi a seguinte:
  - O 1º quartil dos 67 municípios possuía até 5 leitos pediátricos do SUS.
  - O 2º quartil dos 67 municípios possuía até 12 leitos pediátricos do SUS.
  - O 3º quartil dos 67 municípios possuía até 38 leitos pediátricos do SUS.
  - A cidade com maior oferta de leitos pediátricos do SUS foi São Paulo, com 1285 (o outlier no boxplot, que representa 72% da soma de todos os outros 66 municípios), o que denota aumento em relação ao mês anterior.
  - Os 67 municípios paulistas com ao menos 1 caso de COVID-19 no mês 3 possuíam ao todo 3065 leitos pediátricos do SUS.
  - Os altos valores de amplitude (1285), variância (24623) e coeficiente de variação (343%) demonstram como o nº de leitos pediátricos do SUS nos municípios está espalhado e distante da média (45).

Leitos Pediátricos SUS	
Contagem	67
Média	45.74627
Desvio Padrão	156.91769
Mínimo	0
25%	5
50%	12
75%	38.5
Máximo	1285
Soma	3065
Moda	0, 2 e 4
Mediana	12
Amplitude	1285
Variância	24623.16192
Coeficiente de Variação	343.01747



# 4. Análise Exploratória de Dados

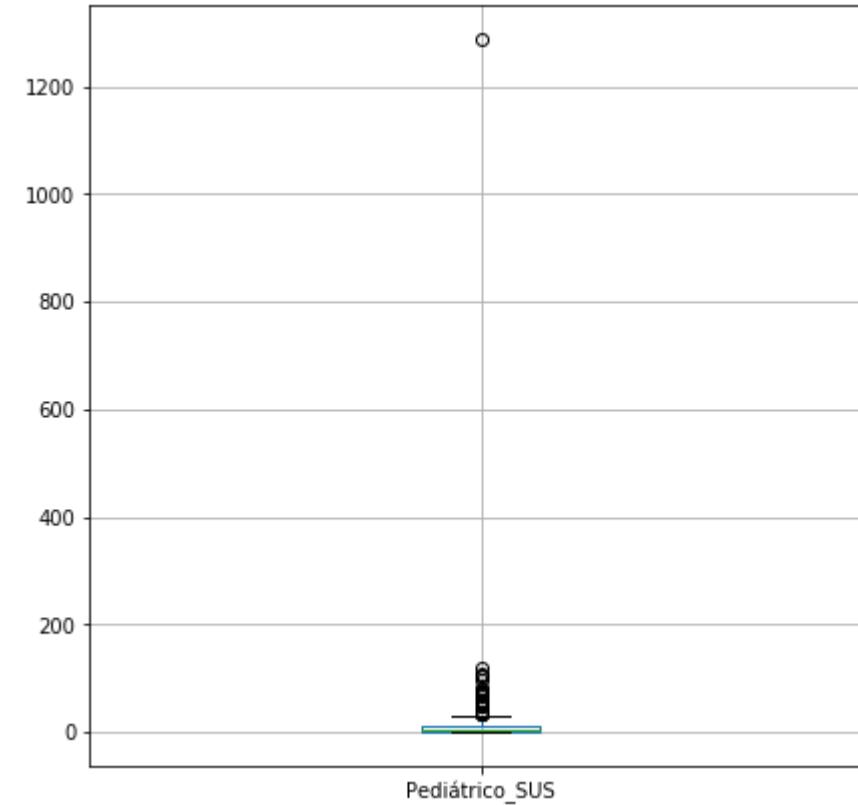
RAIO-X DA BASE | ANÁLISE UNIVARIADA

270

## 'Leitos Pediátricos SUS' – mês 4

- No mês 4, havia 321 municípios do estado de São Paulo com ao menos 1 caso de COVID-19. Destes, a situação de leitos pediátricos do SUS foi a seguinte:
  - O 1º quartil dos 321 municípios não possuía leitos pediátricos do SUS.
  - O 2º quartil dos 321 municípios possuía até 5 leitos pediátricos do SUS.
  - O 3º quartil dos 321 municípios possuía até 12 leitos pediátricos do SUS.
  - A cidade com maior oferta de leitos pediátricos do SUS foi São Paulo, com 1287 (o outlier no boxplot, que representa 37% da soma de todos os outros 320 municípios), o que denota aumento em relação ao mês anterior.
  - Os 321 municípios paulistas com ao menos 1 caso de COVID-19 no mês 4 possuíam ao todo 4743 leitos pediátricos do SUS.
  - Os altos valores de amplitude (1287), variância (5419) e coeficiente de variação (498%) demonstram como o nº de leitos pediátricos do SUS nos municípios está espalhado e distante da média (14).

Leitos Pediátricos SUS	
Contagem	321
Média	14.77570
Desvio Padrão	73.61530
Mínimo	0
25%	0
50%	5
75%	12
Máximo	1287
Soma	4743
Moda	0
Mediana	5
Amplitude	1287
Variância	5419.21203
Coeficiente de Variação	498.21865



# 4. Análise Exploratória de Dados

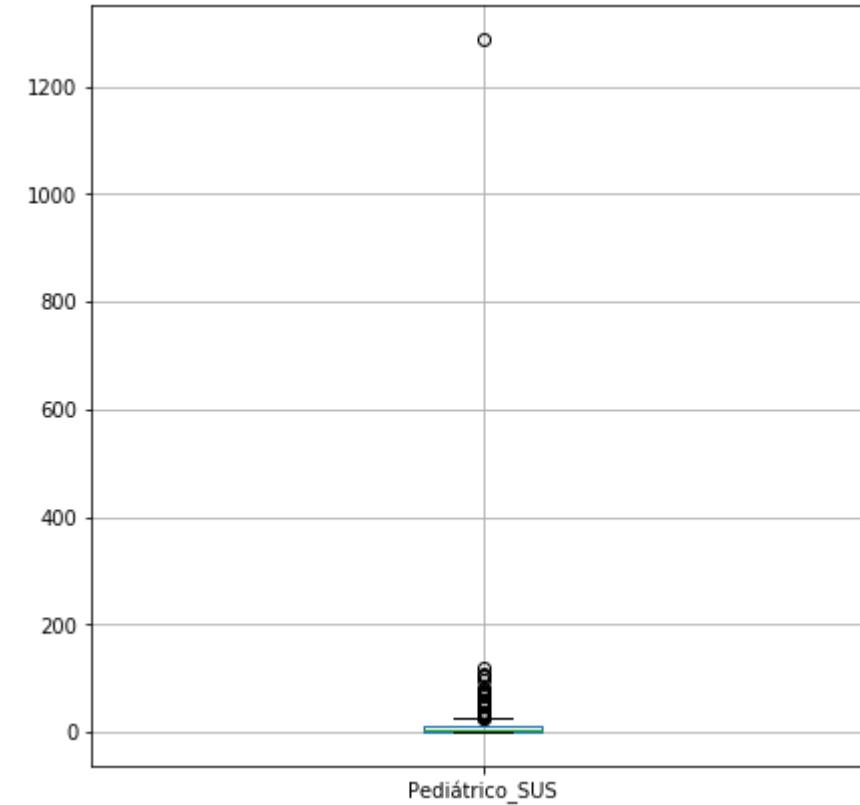
RAIO-X DA BASE | ANÁLISE UNIVARIADA

271

## 'Leitos Pediátricos SUS' – mês 5

- No mês 5, havia 383 municípios do estado de São Paulo com ao menos 1 caso de COVID-19. Destes, a situação de leitos pediátricos do SUS foi a seguinte:
  - O 1º quartil dos 383 municípios não possuía leitos pediátricos do SUS.
  - O 2º quartil dos 383 municípios possuía até 4 leitos pediátricos do SUS.
  - O 3º quartil dos 383 municípios possuía até 10 leitos pediátricos do SUS.
  - A cidade com maior oferta de leitos pediátricos do SUS foi São Paulo, com 1287 (o outlier no boxplot, que representa 35% da soma de todos os outros 382 municípios), ou seja, São Paulo manteve o mesmo nº de leitos pediátricos do SUS em relação ao mês anterior.
  - Os 383 municípios paulistas com ao menos 1 caso de COVID-19 no mês 5 possuíam ao todo 4874 leitos pediátricos do SUS.
  - Os altos valores de amplitude (1287), variância (4562) e coeficiente de variação (530%) demonstram como o nº de leitos pediátricos do SUS nos municípios está espalhado e distante da média (12).

Leitos Pediátricos SUS	
Contagem	383
Média	12.72585
Desvio Padrão	67.54909
Mínimo	0
25%	0
50%	4
75%	10
Máximo	1287
Soma	4874
Moda	0
Mediana	4
Amplitude	1287
Variância	4562.88014
Coeficiente de Variação	530.80228



# 4. Análise Exploratória de Dados

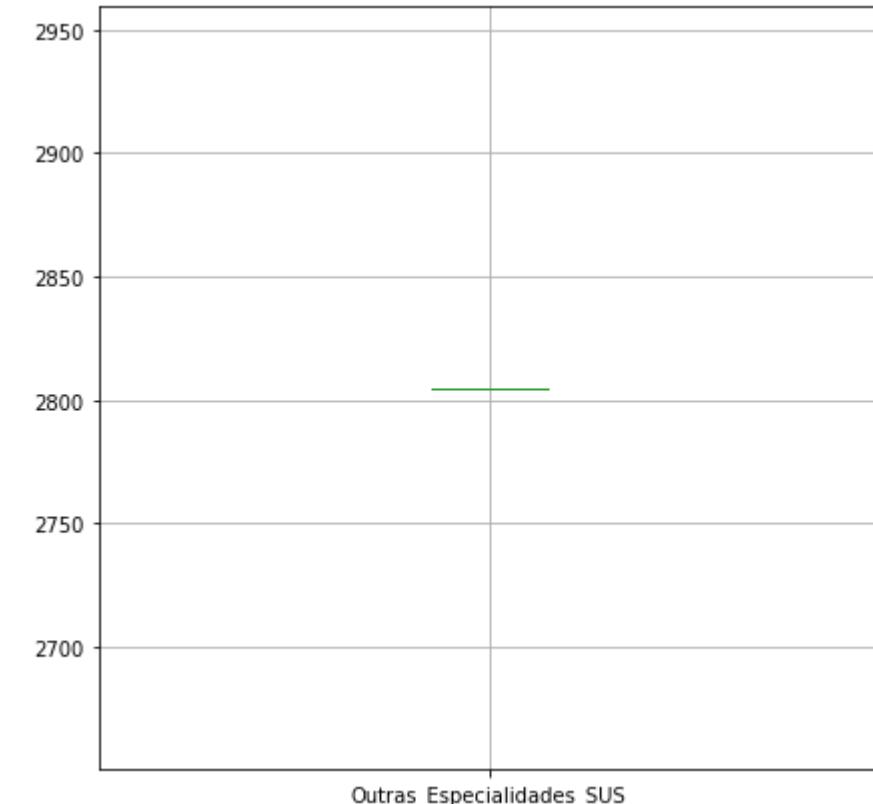
RAIO-X DA BASE | ANÁLISE UNIVARIADA

272

## 'Leitos de Outras Especialidades SUS' – mês 2

- Obs.: não faz sentido fazer análise univariada de nº de leitos hospitalares, pois eles são únicos para cada cidade e para cada mês. Portanto, vamos analisar o nº de leitos por mês.
- No mês 2, só havia observações da cidade de São Paulo, que apresentava 2805 leitos de outras especialidades do SUS.

Leitos de Outras Especialidades SUS	
Contagem	1
Média	2805
Desvio Padrão	-
Mínimo	2805
25%	2805
50%	2805
75%	2805
Máximo	2805
Soma	2805
Moda	2805
Mediana	2805
Amplitude	0
Variância	-
Coeficiente de Variação	-



# 4. Análise Exploratória de Dados

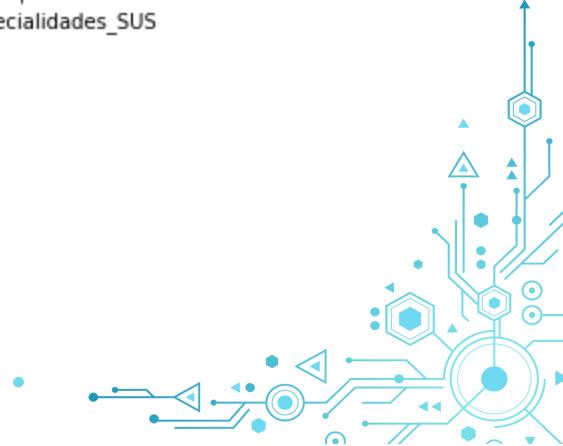
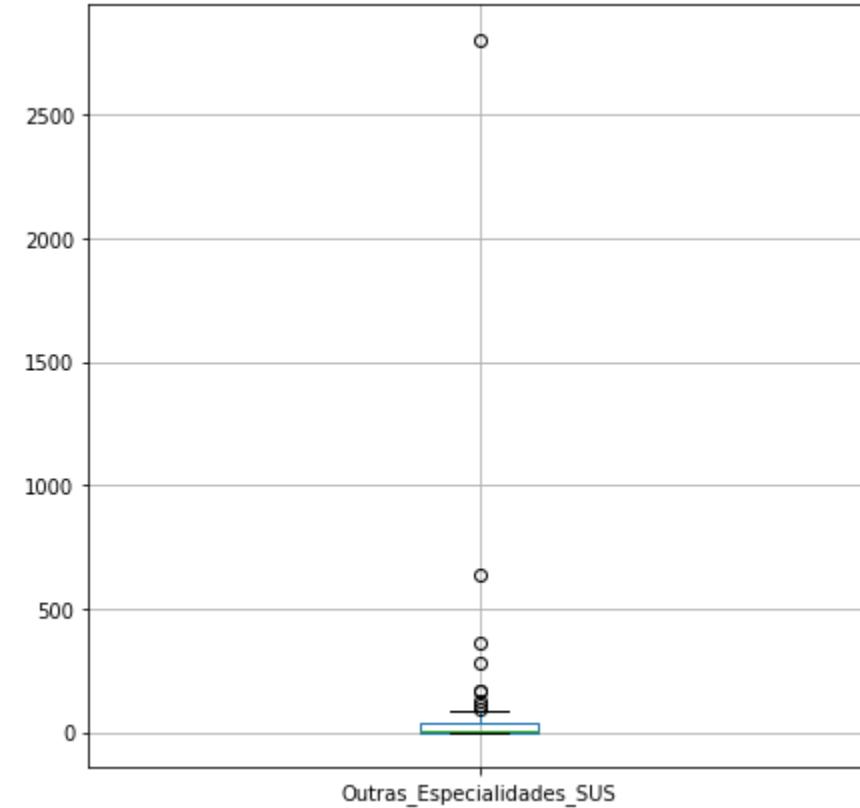
RAIO-X DA BASE | ANÁLISE UNIVARIADA

273

## 'Leitos de Outras Especialidades SUS' – mês 3

- No mês 3, havia 67 municípios do estado de São Paulo com ao menos 1 caso de COVID-19. Destes, a situação de leitos de outras especialidades do SUS foi a seguinte:
  - O 1º quartil dos 67 municípios não possuía leitos de outras especialidades do SUS.
  - O 2º quartil dos 67 municípios possuía até 2 leitos de outras especialidades do SUS.
  - O 3º quartil dos 67 municípios possuía até 37 leitos de outras especialidades do SUS.
  - A cidade com maior oferta de leitos de outras especialidades do SUS foi São Paulo, com 2805 (o outlier no boxplot, que representa mais do que a soma de todos os outros 66 municípios), ou seja, São Paulo manteve o mesmo número de leitos de outras especialidades do SUS do mês anterior.
  - Os 67 municípios paulistas com ao menos 1 caso de COVID-19 no mês 3 possuíam ao todo 5462 leitos de outras especialidades do SUS.
  - Os altos valores de amplitude (2805), variância (123786) e coeficiente de variação (431%) demonstram como o nº de leitos de outras especialidades do SUS nos municípios está espalhado e distante da média (81).

	Leitos de Outras Especialidades SUS
Contagem	67
Média	81.52239
Desvio Padrão	351.83305
Mínimo	0
25%	0
50%	2
75%	37.5
Máximo	2805
Soma	5462
Moda	0
Mediana	2
Amplitude	2805
Variância	123786.49570
Coeficiente de Variação	431.57844



# 4. Análise Exploratória de Dados

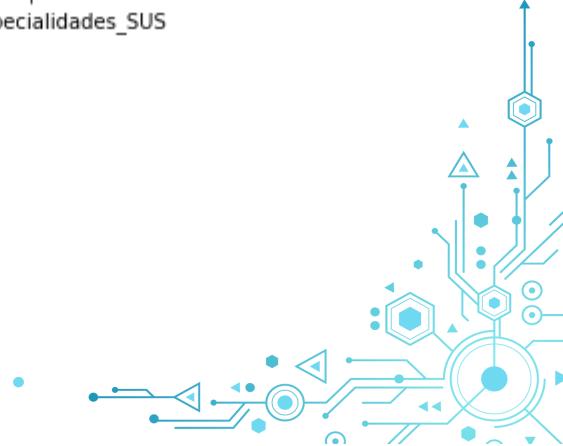
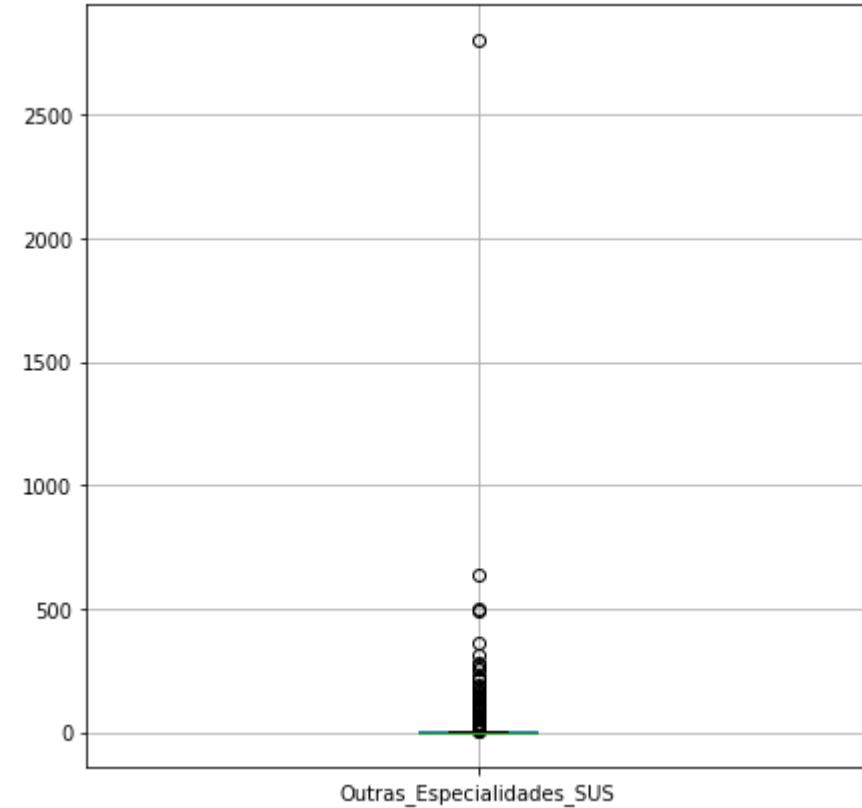
RAIO-X DA BASE | ANÁLISE UNIVARIADA

274

## 'Leitos de Outras Especialidades SUS' – mês 4

- No mês 4, havia 321 municípios do estado de São Paulo com ao menos 1 caso de COVID-19. Destes, a situação de leitos de outras especialidades do SUS foi a seguinte:
  - O 1º e o 2º quartis dos 321 municípios não possuíam leitos de outras especialidades do SUS.
  - O 3º quartil dos 321 municípios possuía até 2 leitos de outras especialidades do SUS.
  - A cidade com maior oferta de leitos de outras especialidades do SUS foi São Paulo, com 2805 (o outlier no boxplot, que representa 37% da soma de todos os outros 320 municípios), ou seja, São Paulo manteve o mesmo nº de leitos de outras especialidades do SUS do mês anterior.
  - Os 321 municípios paulistas com ao menos 1 caso de COVID-19 no mês 4 possuíam ao todo 10377 leitos de outras especialidades do SUS.
  - Os altos valores de amplitude (2805), variância (29705) e coeficiente de variação (533%) demonstram como o nº de leitos de outras especialidades do SUS nos municípios está espalhado e distante da média (32).

Leitos de Outras Especialidades SUS	
Contagem	321
Média	32.32710
Desvio Padrão	172.35250
Mínimo	0
25%	0
50%	0
75%	2
Máximo	2805
Soma	10377
Moda	0
Mediana	0
Amplitude	2805
Variância	29705.38329
Coeficiente de Variação	533.15170



## 4. Análise Exploratória de Dados

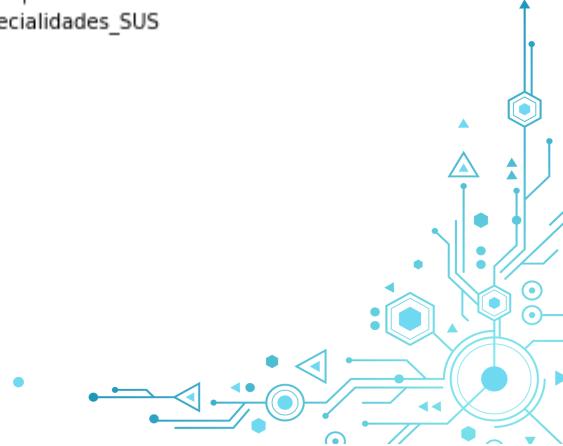
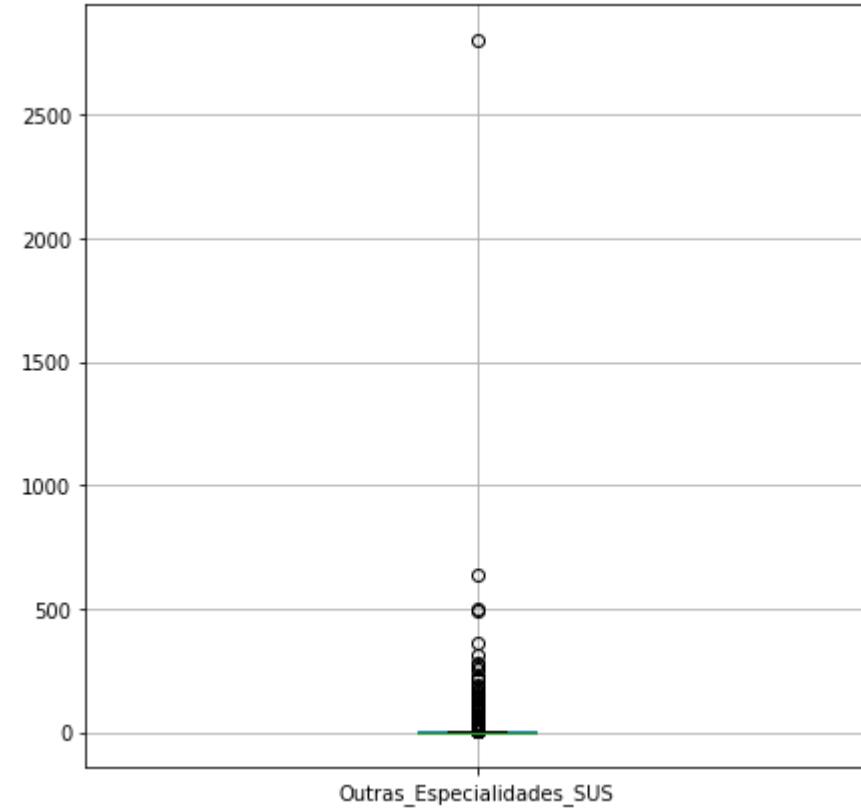
RAIO-X DA BASE | ANÁLISE UNIVARIADA

275

### 'Leitos de Outras Especialidades SUS' – mês 5

- No mês 5, havia 383 municípios do estado de São Paulo com ao menos 1 caso de COVID-19. Destes, a situação de leitos de outras especialidades do SUS foi a seguinte:
  - O 1º e o 2º quartis dos 383 municípios não possuíam leitos de outras especialidades do SUS.
  - O 3º quartil dos 383 municípios possuía até 1 leito de outras especialidades do SUS.
  - A cidade com maior oferta de leitos de outras especialidades do SUS foi São Paulo, com 2805 (o outlier no boxplot, que representa 37% da soma de todos os outros 382 municípios), mantendo o mesmo nº do mês anterior.
  - Os 383 municípios paulistas com ao menos 1 caso de COVID-19 no mês 5 possuíam ao todo 10383 leitos de outras especialidades do SUS.
  - Os altos valores de amplitude (2805), variância (25025) e coeficiente de variação (583%) demonstram como o nº de leitos de outras especialidades do SUS nos municípios está espalhado e distante da média (27).

Leitos de Outras Especialidades SUS	
Contagem	383
Média	27.10966
Desvio Padrão	158.19434
Mínimo	0
25%	0
50%	0
75%	1
Máximo	2805
Soma	10383
Moda	0
Mediana	0
Amplitude	2805
Variância	25025.44868
Coeficiente de Variação	583.53493



# 4. Análise Exploratória de Dados

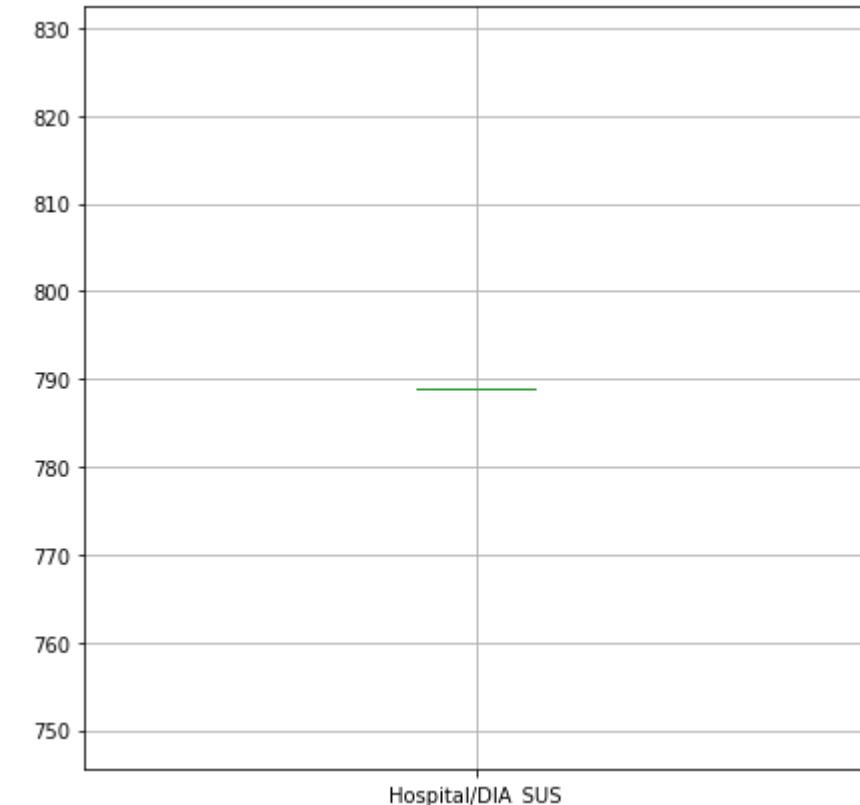
RAIO-X DA BASE | ANÁLISE UNIVARIADA

276

## 'Leitos Hospital/Dia SUS' – mês 2

- Obs.: não faz sentido fazer análise univariada de nº de leitos hospitalares, pois eles são únicos para cada cidade e para cada mês. Portanto, vamos analisar o nº de leitos por mês.
- No mês 2, só havia observações da cidade de São Paulo, que apresentava 789 leitos Hospital/DIA do SUS.

	Leitos Hospital/Dia SUS
Contagem	1
Média	789
Desvio Padrão	-
Mínimo	789
25%	789
50%	789
75%	789
Máximo	789
Soma	789
Moda	789
Mediana	789
Amplitude	0
Variância	-
Coeficiente de Variação	-



## 4. Análise Exploratória de Dados

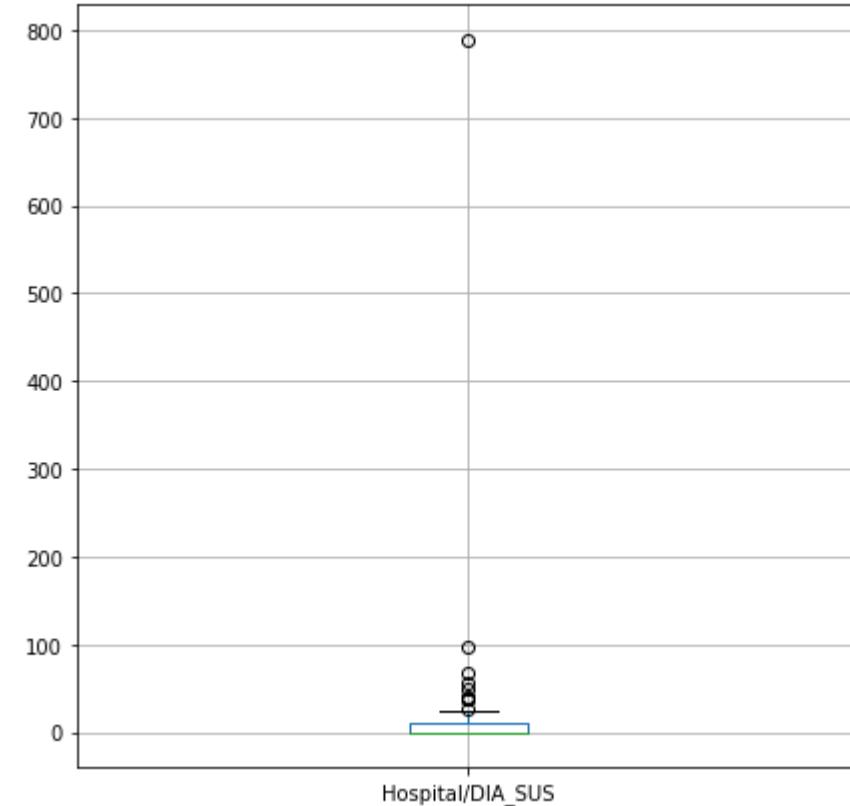
RAIO-X DA BASE | ANÁLISE UNIVARIADA

277

### 'Leitos Hospital/Dia SUS' – mês 3

- No mês 3, havia 67 municípios do estado de São Paulo com ao menos 1 caso de COVID-19. Destes, a situação de leitos Hospital/DIA do SUS foi a seguinte:
  - O 1º e o 2º quartis dos 67 municípios não possuíam leitos Hospital/DIA do SUS.
  - O 3º quartil dos 67 municípios possuía até 10 leitos Hospital/DIA do SUS.
  - A cidade com maior oferta de leitos Hospital/DIA do SUS foi São Paulo, com 789 (o outlier no boxplot, que representa 25% a mais do que a soma de todos os outros 66 municípios), mantendo o mesmo nº do mês anterior.
  - Os 67 municípios paulistas com ao menos 1 caso de COVID-19 no mês 3 possuíam ao todo 1416 leitos Hospital/DIA do SUS.
  - Os altos valores de amplitude (789), variância (9411) e coeficiente de variação (459%) demonstram como o nº de leitos Hospital/DIA do SUS nos municípios está espalhado e distante da média (21).

	Leitos Hospital/Dia SUS
Contagem	67
Média	21.13433
Desvio Padrão	97.01185
Mínimo	0
25%	0
50%	0
75%	10.5
Máximo	789
Soma	1416
Moda	0
Mediana	0
Amplitude	789
Variância	9411.29986
Coeficiente de Variação	459.02502



# 4. Análise Exploratória de Dados

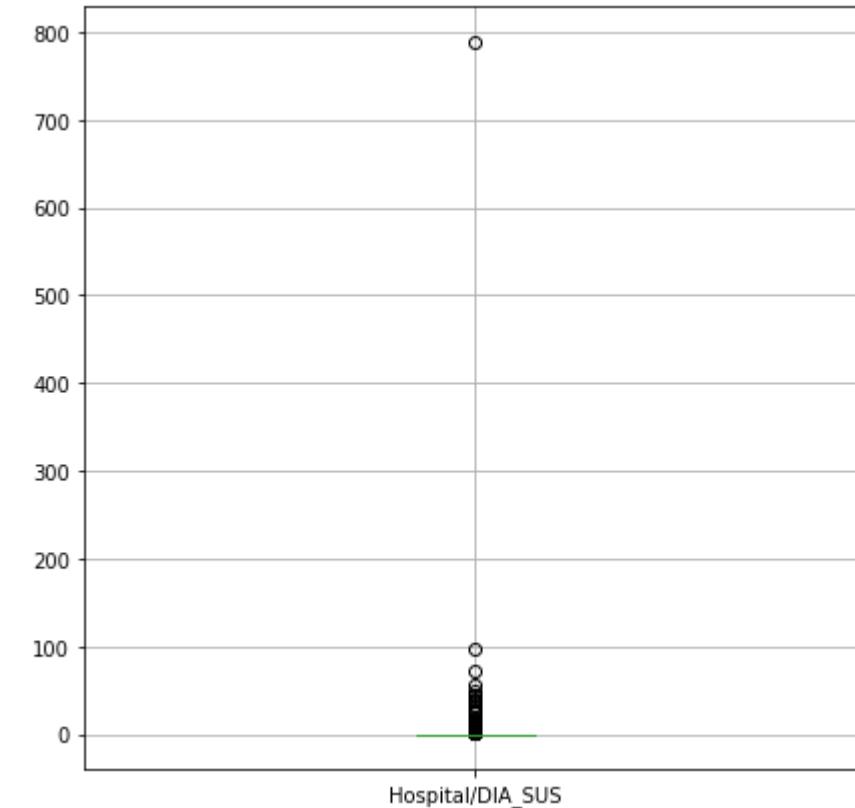
RAIO-X DA BASE | ANÁLISE UNIVARIADA

278

## 'Leitos Hospital/Dia SUS' – mês 4

- No mês 4, havia 321 municípios do estado de São Paulo com ao menos 1 caso de COVID-19. Destes, a situação de leitos Hospital/DIA do SUS foi a seguinte:
- O 1º, o 2º e o 3º quartis dos 321 municípios não possuíam leitos Hospital/DIA do SUS.
- A cidade com maior oferta de leitos Hospital/DIA do SUS foi São Paulo, com 789 (o outlier no boxplot, que representa 82% da soma de todos os outros 320 municípios), mantendo o mesmo nº do mês anterior.
- Os 321 municípios paulistas com ao menos 1 caso de COVID-19 no mês 4 possuíam ao todo 1747 leitos Hospital/DIA do SUS.
- Os altos valores de amplitude (789), variância (2030) e coeficiente de variação (827%) demonstram como o nº de leitos Hospital/DIA do SUS nos municípios está espalhado e distante da média (5).

	Leitos Hospital/Dia SUS
Contagem	321
Média	5.44237
Desvio Padrão	45.06083
Mínimo	0
25%	0
50%	0
75%	0
Máximo	789
Soma	1747
Moda	0
Mediana	0
Amplitude	789
Variância	2030.47870
Coeficiente de Variação	827.96379



# 4. Análise Exploratória de Dados

RAIO-X DA BASE | ANÁLISE UNIVARIADA

279

## 'Leitos Hospital/Dia SUS' – mês 5

- No mês 5, havia 383 municípios do estado de São Paulo com ao menos 1 caso de COVID-19. Destes, a situação de leitos Hospital/DIA do SUS foi a seguinte:
  - O 1º, o 2º e o 3º quartis dos 383 municípios não possuía leitos Hospital/DIA do SUS.
  - A cidade com maior oferta de leitos Hospital/DIA do SUS foi São Paulo, com 789 (o outlier no boxplot, que representa 82% da soma de todos os outros 382 municípios), ou seja, São Paulo manteve o mesmo nº de leitos Hospital/DIA do SUS em relação ao mês anterior.
  - Os 383 municípios paulistas com ao menos 1 caso de COVID-19 no mês 5 possuíam ao todo 1747 leitos Hospital/DIA do SUS.
  - Os altos valores de amplitude (789), variância (1704) e coeficiente de variação (905%) demonstram como o nº de leitos Hospital/DIA do SUS nos municípios está espalhado e distante da média (4).

	Leitos Hospital/Dia SUS
Contagem	383
Média	4.56136
Desvio Padrão	41.29108
Mínimo	0
25%	0
50%	0
75%	0
Máximo	789
Soma	1747
Moda	0
Mediana	0
Amplitude	789
Variância	1704.95369
Coeficiente de Variação	905.23672

